14:12:49 OCA PAD INITIATION - PROJECT HEADER INFORMATION 04/19/88

Active

Project #: E-24-612 Cost share #: Rev #: 0
Center # : R6461-0A0 Center shr #: OCA file #:
Work type : RES
Contract#: P-7917(8899)-1108 Mod #: Document : SUBCONT
Prime #: DLA900-86-C-2045 Contract entity: GTRC

Subprojects ? : N
Main project #:

Project unit: ISYE Unit code: 02.010.124
Project director(s):
BANKS J ISYE

Sponsor/division names: BATTELLE /
Sponsor/division codes: 500 / 007

Award period: 880201 to 890131 (performance) 890131 (reports)

Sponsor amount	New this change	Total to date
Contract value	40,392.00	40,392.00
Funded	40,392.00	40,392.00
Cost sharing amount		0.00

Does subcontracting plan apply ?: N

Title: TESTING, UNDERSTANDING AND VALIDATING COMPLEX SIMULATION MODELS

PROJECT ADMINISTRATION DATA

OCA contact: Ina R. Lashley 894-4820

Sponsor technical contact

MS KAREN SISTEK
(614)424-6424
BATTELLE COLUMBUS DIVISION
505 KING AVE
COLUMBUS OHIO 43201-2693

Sponsor issuing office

MR ROBERT E TANNER
(614)424-4521
BATTELLE COLUMBUS DIVISION
505 KING AVE
COLUMBUS OHIO 43201-2693

Security class (U,C,S,TS) : U
Defense priority rating : DO-A7
Equipment title vests with: Sponsor
NONE PROPOSED.

ONR resident rep. is ACO (Y/N): N
N/A supplemental sheet
GIT X

Administrative comments -
PROJECT INITIATION. FOREIGN NATIONALS RESTRICTED (ART. XIX). SUBCONTRACTS REQUIRE PRIOR APPROVAL. LTR DTD 24 FEB 1988 IS A PART OF THIS SUBCONTRACT

GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION

SR492

NOTICE OF PROJECT CLOSEOUT

Date 3/8/89

roject No. E-24-612 Center No. R6461-0A0

roject Director J. Banks School/Lab ISyE

ponsor Battelle

ontract/Grant No. P7917(8899)-1108 GTRC XX GIT ____

ime Contract No. DLA900-86-C-2045

tle Testing, Understanding, and Validating Complex Simulation Models

fective Completion Date 4/30/89 (Performance) 4/30/89 (Reports)

oseout Actions Required:

__ None
__ Final Invoice or Copy of Last Invoice
__ Final Report of Inventions and/or Subcontracts- Patent Questionnaire sent to PI.
__ Government Property Inventory & Related Certificate
__ Classified Material Certificate
__ Release and Assignment
__ Other _____

cludes Subproject No(s). _____

bproject Under Main Project No. _____

ntinues Project No. _____ Continued by Project No. _____

stribution:

_ Project Director
_ Administrative Network
_ Accounting
_ Procurement/GTRI Supply Services
_ Research Property Management
_ Research Security Services

X Reports Coordinator (OCA)
X GTRC
X Project File
X Contract Support Division (OCA)(2)
__ Other _____

E-24-612

TESTING, UNDERSTANDING AND VALIDATING COMPLEX SIMULATION MODELS

Prime Contract No. DLA900-86-C-2045
BCD Subcontract No. P-7917(8899)-1108
GT Research Project No. E-24-612

First Quarterly Report

February 17, 1988-April 30, 1988

Prepared for
Batelle
Edgewood Operations
2113 Emmorton Park Road
Edgewood, MD 21040

Prepared by
Jerry Banks, Principal Investigator
Georgia Institute of Technology
Atlanta, GA 30332

(Attention: Nancy Brletich)

I. Technical

Task 1. Development of Analogies from Testing, Understanding and Validating Physical Systems

A. Effort Accomplished

i. Developed eleven analogies with inferences for complex military systems.
ii. Summarized inferences.

B. Changes from the proposed work

i. Analogies were extended from physical systems to include service systems since physical systems are an artificial boundary. Awaiting response from client.
ii. An Army student assigned this task withdrew from the project at the mid-point of the reporting period. Task was undertaken by the Principal Investigator.

C. New directions from client

None

D. Deliverables submitted

None

Task 2. Use of Statistical Models

A. Effort accomplished

i. Selected four statistical areas for application.
ii. Prepared descriptions of the application of the statistical methods in each area to complex simulation models.
iii. Described how the statistical methods apply to ATCAL.

B. Changes from the proposed work

i. One of the two Army students assigned this task moved to Task III at the mid-point of the reporting period. Second Army student accepted responsibility for the task.

C. New direction from the client

None

D. Deliverables submitted

None

Task 3. Extension of Validation Methods

A. Effort accomplished

i. Collected vast amounts of information on verification and validation methods.
ii. Examined several innovative verification and validation examples which have application to this task.
iii. Analyzed existing verification and validation methods and determined those which can be used to evaluate complex models, those which can be used with some modification, and those which appear not to be useful.

B. Changes from the proposed work

i. Army student assigned this task withdrew at the mid-point of the reporting period. Another of the Army students on the project accepted responsibility for this task.

C. New directions from client

None

D. Deliverables submitted

None

II. Trips Taken

U.S. Army Concepts Analysis Agency
Bethesda, MD
March 30, 1988

U.S. Army TRAC-FLVN
Ft. Leavenworth, KS
April 15, 1988

U.S. Army TRAC-MTRY
Monterey, CA
April 19, 1988

III. Cost

See attached pages

PRINCIPAL INVESTIGATOR J BANKS CENTER NO. 243R64610AO ACCOUNT NO. E-24-612

STATUS AT END OF APRIL 1988 DEPARTMENT I & S ENG

SPONSOR BATTELLE GTRC

AWARD NUMBER P-7917(8899)-1108 RF CENTER NO. 003460100000 RESTRICTED FUND RF-49085

EFFECTIVE DATE 02-01-88 BILLING GROUP GTRC EXPIRATION DATE 01-31-89

	MONTH	FISCAL YEAR	TOTAL CONTRACT	
OVERHEAD				
BUDGET			15,147.00	
EXPENDED	438.66	439.14	439.14	RATE OF 6.0 BASE OF 4
ENCUMBERED	200.81	271.31	271.31	
FREE BALANCE			14,436.55	
TOTAL				
BUDGET			40,392.00	
EXPENDED	7,749.60	7,758.12	7,758.12	
ENCUMBERED	3,547.57	4,793.07	4,793.07	
FREE BALANCE			27,840.81	

GEORGIA INSTITUTE OF TECHNOLOGY

PRINCIPAL INVESTIGATOR J BANKS CENTER NO. 243R64610AO ACCOUNT NO. E-24-612

STATUS AT END OF APRIL 1988 DEPARTMENT I & S ENG

SPONSOR BATTELLE GTRC

AWARD NUMBER P-7917(8899)-1108 RF CENTER NO. 00346010000 RESTRICTED FUND RF-49085

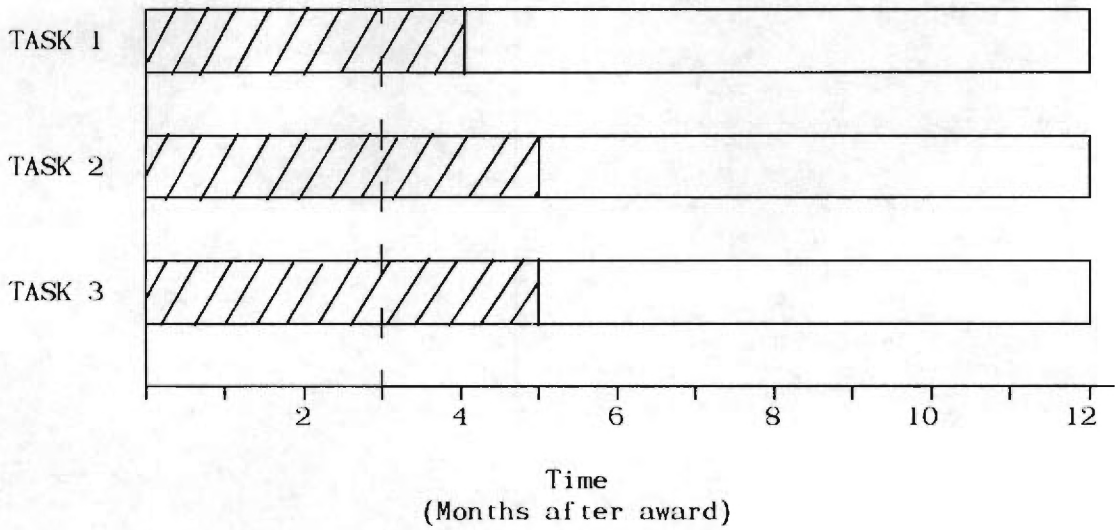
EFFECTIVE DATE 02-01-88 BILLING GROUP GTRC EXPIRATION DATE 01-31-89

	MONTH	FISCAL YEAR	TOTAL CONTRACT
PERSONAL SERVICES			
BUDGET			16,650.00
EXPENDED	5,550.00	5,550.00	5,550.00
ENCUMBERED	2,775.00	2,775.00	2,775.00
FREE BALANCE			8,325.00
FRINGE BENEFITS			
BUDGET			4,595.00
EXPENDED	1,531.80	1,531.80	1,531.80
ENCUMBERED	765.90	765.90	765.90
FREE BALANCE			2,297.30
MATERIALS AND SUPPLIES			
BUDGET			1,000.00
EXPENDED	.00	.00	.00
ENCUMBERED	.00	.00	.00
FREE BALANCE			1,000.00
TRAVEL			
BUDGET			3,000.00
EXPENDED	229.14	237.18	237.18
ENCUMBERED	-194.14	980.86	980.86
FREE BALANCE			1,781.96
TOTAL DIRECT CHARGES			
BUDGET			25,245.00
EXPENDED	7,310.94	7,318.98	7,318.98
ENCUMBERED	3,346.76	4,521.76	4,521.76
FREE BALANCE			13,404.26

IV. Problem Areas

None

PROGRAM STATUS



E-24-612

TESTING, UNDERSTANDING, AND VALIDATING COMPLEX SIMULATION MODELS

Prime Contract No. DLA900-86-C-2045
BCD Subcontract No. P-7917(8899)-1108
GT Research Project No. E-24-612

Second Quarterly Report

May 1, 1988 - July 31, 1988

Prepared for
Batelle
Edgewood Operations
2113 Emmorton Park Road
Edgewood, MD 21040

Prepared by
Jerry Banks, Principal Investigator
Georgia Institute of Technology
Atlanta, GA 30332

(Attention: Nancy Brletich)

I. Technical

Task 1. Development of Analogies from Testing, Understanding and Validating Physical Systems

A. Effort accomplished

i. Preparation of Draft Final Report.

ii. Contacts made with NASA-MSFC, CACI, and FAA to discuss verification of large scale computer programs.

iii. Seeking additional contacts with FAA and NRC.

B. Changes from the proposed work

None

C. New directions from client

Extend analogies to verification of large scale computer programs developed by other agencies.

D. Deliverables submitted

Analogies and inferences.
Included in Draft Final Report submitted to sponsor on June 10, 1988 and briefed on July 22, 1988.

Task 2. Use of Statistical Models

A. Effort accomplished

i. Begun consideration of application of regression models.

ii. Preparation of Draft Final Report.

B. Changes from the proposed work

None

C. New direction from the client

Stand alone software or module from large program may be sent for example applications of suggested methods.

D. Deliverables submitted

Four statistical methods with applications to ATCAL included in Draft Final Report submitted to sponsor on June 10, 1988 and briefed on July 22, 1988.

Task 3. Extension of Validation Methods

A. Effort accomplished

Preparation of Draft Final Report

B. Changes from the proposed work

None

C. New directions from client

Contact LTC Vern Betancourt for second visit to TRAC-MTRY.

D. Deliverables submitted

Extensions included in Draft Final Report submitted to sponsor on June 10, 1988 and briefed on July 22, 1988.

II. Trips Taken

U.S. Army Concepts Analysis Agency
Bethesda, MD
July 22, 1988

III. Cost

See attached page, Note, July status report not available until August 9, 1988.

IV. Problem Areas

None

PRINCIPAL INVESTIGATOR J BANKS CENTER NO. 243R64610AO ACCOUNT NO. E-24-612

STATUS AT END OF JUNE 1988 DEPARTMENT I & S ENG

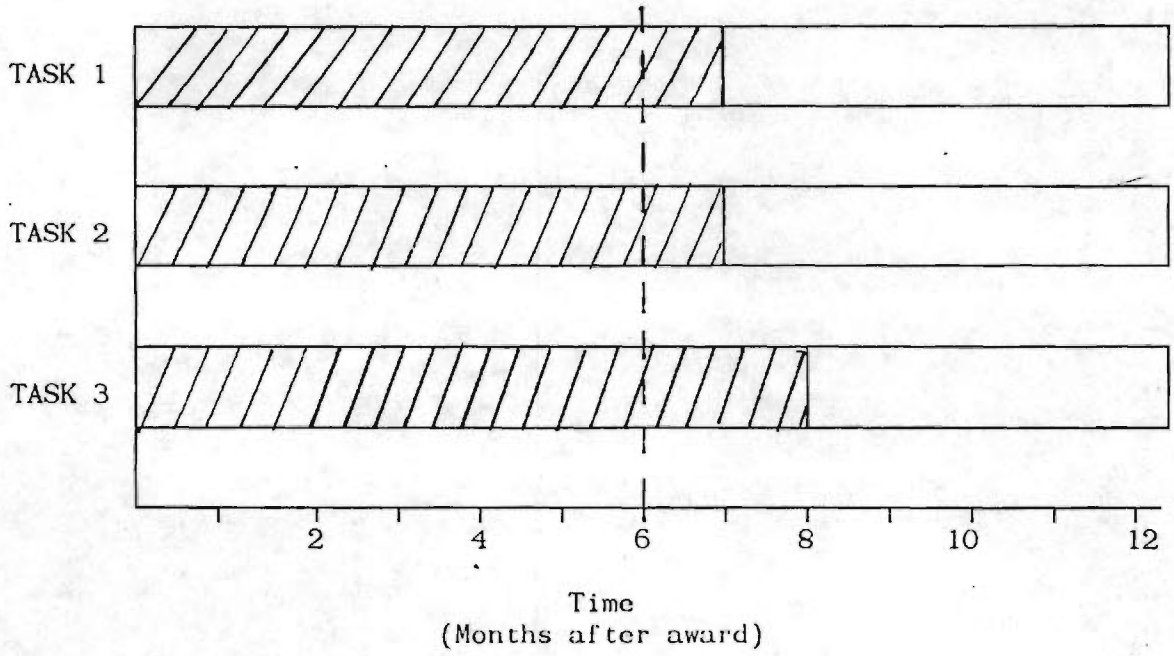
SPONSOR BATTELLE GTRC

AWARD NUMBER P-7917(8899)-1108 RF CENTER NO. 00346010000 RESTRICTED FUND RF-49085

EFFECTIVE DATE 02-01-88 BILLING GROUP GTRC EXPIRATION DATE 01-31-89

	MONTH	FISCAL YEAR	TOTAL CONTRACT
PERSONAL SERVICES			
BUDGET			16,650.00
EXPENDED	1,387.50	8,325.00	8,325.00
ENCUMBERED	-1,387.50	.00	.00
FREE BALANCE			8,325.00
FRINGE BENEFITS			
BUDGET			4,595.00
EXPENDED	382.95	2,297.70	2,297.70
ENCUMBERED	-382.95	.00	.00
FREE BALANCE			2,297.30
MATERIALS AND SUPPLIES			
BUDGET			1,000.00
EXPENDED	.00	.00	.00
ENCUMBERED	.00	.00	.00
FREE BALANCE			1,000.00
TRAVEL			
BUDGET			3,000.00
EXPENDED	296.04	1,200.56	1,200.56
ENCUMBERED	-35.00	.00	.00
FREE BALANCE			1,799.44
TOTAL DIRECT CHARGES			
BUDGET			25,245.00
EXPENDED	2,066.49	11,823.26	11,823.26
ENCUMBERED	-1,805.45	.00	.00
FREE BALANCE			13,421.74

PROGRAM STATUS



E-24-612

TESTING, UNDERSTANDING, AND VALIDATING COMPLEX SIMULATION MODELS

Prime Contract No. DLA900-86-C-2045
BCD Subcontract No. P-7917(8899)-1108
GT Research Project No. E-24-612

Third Quarterly Report

August 1, 1988 - October 31, 1988

Prepared for
Batelle
Edgewood Operations
2113 Emmorton Park Road
Edgewood, MD 21040

Prepared by
Jerry Banks, Principal Investigator
Georgia Institute of Technology
Atlanta, GA 30332

(Attention: Nancy Brletich)

I. Technical

Task 1. Development of Analogies from Testing, Understanding and Validating Physical Systems

A. Effort accomplished

i. Visited with MITRE Corporation to discuss FAA activity.

ii. Obtained information from MITRE concerning SDI verification and validation.

iii. Additional contacts with Hartsfield Airport, Plant Hatch and Southern Railway.

B. Changes from the proposed work

None

Task 2. Use of Statistical Models

A. Effort accomplished

i. Use of regression modeling was terminated after investigation and discussion with sponsor.

ii. Considering application of time series models.

iii. Investigated TRANSMO as a possible software example for application of statistical methods.

B. Changes from the proposed work

None

C. New direction from the client

Discontinue concentration on TRANSMO.

Task 3. Extension of Validation Methods

A. Effort accomplished

i. Contact with Prof. Ingber from NPGS and receipt of information on NTC.

ii. Contact with MAJ Galing with promise of a copy of his dissertation which describes Turing test for NTC data.

iii. Receipt of information from Dr. Brown, Auburn University, concerning automated V/V.

iv. Contact with COL Evans, AMIP.

B. Changes from the proposed work

None

II. Trips Taken

Visit to MITRE and
Presentation to Mr. Walter Hollis, DUSA-OR
Washington, DC Area
September 27, 1988
Jerry Banks

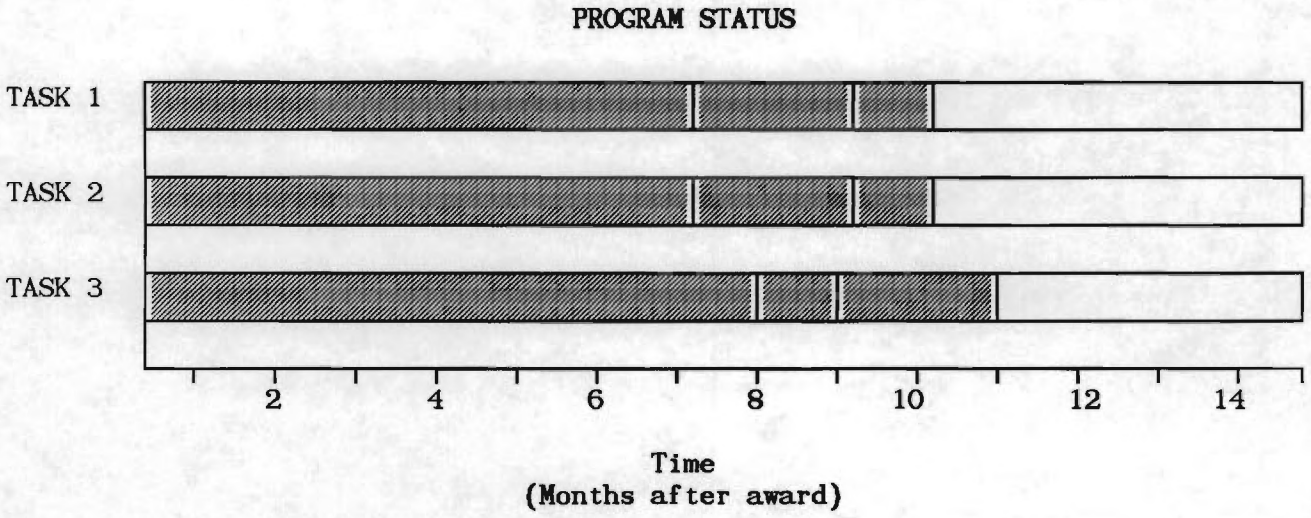
U.S. Army Concepts Analysis Agency
Bethesda, MD
November 1, 1988
CPTs Dawson and Scott

III. Cost

See attached page.

IV. Problem Areas

None


PROGRAM STATUS
TASK 1
TASK 2
TASK 3
2
4
6
8
10
12
14
Time
(Months after award)

PRINCIPAL INVESTIGATOR J BANKS CENTER NO. 246R64610AO ACCOUNT NO. E-24-612

STATUS AT END OF SEPTEMBER 1988 DEPARTMENT I & S ENG

SPONSOR BATTELLE GTRC

AWARD NUMBER P-7917(8899)-1108 RF CENTER NO. 00646010000 RESTRICTED FUND RF-49085

EFFECTIVE DATE 02-01-88 BILLING GROUP GTRC EXPIRATION DATE 01-31-89

	MONTH	FISCAL YEAR	TOTAL CONTRACT
PERSONAL SERVICES			
BUDGET			16,650.00
EXPENDED	1,389.50	4,168.50	12,493.50
ENCUMBERED	-1,389.50	4,119.99	4,119.99
FREE BALANCE			36.51
FRINGE BENEFITS			
BUDGET			4,595.00
EXPENDED	354.32	1,062.96	3,360.66
ENCUMBERED	-354.32	1,050.60	1,050.60
FREE BALANCE			183.74
MATERIALS AND SUPPLIES			
BUDGET			1,000.00
EXPENDED	.00	36.50	36.50
ENCUMBERED	50.00	50.00	50.00
FREE BALANCE			913.50
TRAVEL			
BUDGET			3,000.00
EXPENDED	544.00	677.73	1,878.29
ENCUMBERED	-545.27	30.00	30.00
FREE BALANCE			1,091.71
TOTAL DIRECT CHARGES			
BUDGET			25,245.00
EXPENDED	2,287.82	5,945.69	17,768.95
ENCUMBERED	-2,239.09	5,250.59	5,250.59
FREE BALANCE			2,225.46

GEORGIA TECH 1885-1985

DESIGNING TOMORROW TODAY

Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia 30332-0205
(404) 894-2300

February 17, 1989

Mr. Howard Whitley
US Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814-2797

Dear Howard:

Inclosed you will find a copy of the Final Report entitled "Testing, Understanding, and Validating Complex Simulation Models". I am sending this copy unbound so that you can make additional copies as desired by CAA. The formal copy of the Final Report will make its way to you through the official channels, but that can take a few weeks.

I believe that the Final Report has answered all of the questions and responded to all of the suggestions that were made in our meeting of meeting of July 22, 1988, and subsequent conversations.

It was a pleasure working with you, Mr. Vandiver, and Gerry Cooper on this project. I also appreciate the help given by many others at CAA, Carl Bates, Dan Shedlowski, and Bob McQuie, to mention a few.

For reference purposes, our internal project number is E-24-612. The Contract number is P-7917(8899)-1108. The prime contract number is DLA900-86-C-2045.

Best regards,

Jerry Banks

Incl.

cc: Fran Cochran, ISyE ✓

Final Report

Testing, Understanding and Validating Complex Simulation Models

Jerry Banks
Georgia Institute of Technology

CPT Mike Casas
US Army Concepts Analysis Agency

CPT Jim Boatner
US Air Force Academy

CPT John Scott
Georgia Institute of Technology

CPT Donald Dawson
Georgia Institute of Technology

Prepared for
US Army Concepts Analysis Agency

February 1989

Table of Contents

Executive Summary

Testing, Understanding, and Validating Complex Simulation Models

Jerry Banks
CPT Mike Casas
CPT Jim Boatner
CPT Donald Dawson
CPT John Scott

Georgia Institute of Technology

This research prescribes new methods for testing, understanding, and validating complex simulation models currently in use by CAA. The new methodology was developed by drawing analogies from systems in general, drawing inferences from specific systems, using statistical methods in new ways, and by extending current validation methods.

The research for this project included an extensive review of existing verification and validation literature in the civilian and military communities. Numerous site visits and telephone calls were made to gain insight into the subject.

This document is organized into eight sections. Section I identifies the purpose for this research and discusses the background leading to this project. Section II provides a definition of terms. Section III discusses the conduct of the research. The bulk of the research results are contained in sections IV through VII and focus primarily on four subject areas:

1. Analogies from testing, understanding, and validating systems in general (Section IV). These analogies are drawn from an examination of

banking, physical examinations (of the human anatomy), medicine and drugs, restaurants, academic promotion and tenure, business, religion, election for political office, building construction, medical diagnosis, the Ph.D. program, food and nutrition, the airline industry, personal transportation, fire fighting, and marketing. All of these systems have characteristics which define them and procedures for maintaining high standards of performance, accountability, and reliability. These various characteristics, procedures, and techniques are examined and analogies are drawn from them for applicability to military modeling. A comprehensive list of the analogies is provided at the end of Section IV.

2. Inferences drawn from large specific systems (Section V).

Large, but specific, systems are examined to identify those attributes and characteristics which make them work well on a daily basis. The TAC Thunder wargame, the Strategic Defense System, the National Airspace System, the Marshall Space Flight Center, and the Plant Hatch nuclear reactor are examined in this study. By examining these systems, a number of validation and verification strategies are drawn into a systematic proposal for the CAA analyst. This proposal includes an initial problem review, model requirements review, review of simulation development, conceptual model assessment, software verification, determination of operational validity, and determination of data validity for the CAA model. This methodology and thought process is provided at the end of the section.

3. Use of statistical methods (Section VI). Control charts, acceptance sampling, fractional factorial analysis, cluster analysis, regression analysis, and time-series analysis techniques are examined for applicability to military models. Applications of these techniques are applied to the military models ATCAL and TRANSMO to demonstrate possible utilization in a military context. Acceptance sampling, fractional factorial analysis, and cluster analysis methods are cost effective techniques for the CAA analyst working under resource constraints.

4. Extension of current validation methods (Section VII). This portion of the research examines the components and concepts of verification, validation, and model credibility as they are traditionally used in the modeling community. It also identifies other factors such as the quality of input data, initial planning of the review process, the purpose of a model, configuration control, and documentation, and assesses their impact on model credibility. The applicability of traditional verification/validation techniques is discussed and extended to CAA's stated goals for establishing model credibility. Innovative military applications of traditional verification/validation techniques are identified for possible adoption by CAA. This research is presented in section VII.

Section VIII contains recommendations to CAA for the application of this research. These recommendations are a synthesis of the research presented in Sections IV through VII. Examples of the recommendations are as follows: CAA must review and determine the current utility and

future worth of a model prior to initiating verification/validation. The specific goals and objectives of the process must then be specified. Section V provides a seventeen step process to which CAA should adhere when planning and executing the verification/validation of a particular model. The techniques of control charting, acceptance sampling, fractional factorial experimental design, and cluster analysis are particularly useful in a resource constrained environment (Section VI). Automated input data review should be developed at CAA to reduce the time and expense of validating input data bases (section VII). Additionally, the incorporation of operational graphics into a model enhances user interaction and understanding during model execution.

It is unlikely that time, personnel, hardware, and software available would enable CAA to apply all of the verification/validation techniques in this document to a particular model. However, careful determination of the credibility needs of a particular model, cross-referenced against the techniques provided in this document, should enable the CAA analyst to determine a verification/validation methodology that is both cost effective and meaningful.

I. Introduction

A. Purpose.

This research proposes new methodologies for testing, understanding, and increasing the credibility of large, complex simulation models used by the Concepts Analysis Agency (CAA). The research focuses on four areas of investigation:

1. Analogies from testing, understanding, and validating systems in general.
2. Inferences drawn from specific large systems.
3. Use of statistical methods.
4. Extension of current validation methods.

The analogies are drawn from diverse systems such as banking, medical diagnosis, and the promotion and tenure system in place at a typical university. Analogies are drawn from these systems to the verification and validation of military simulation models developed by CAA. Traditional statistical methods such as control charting, acceptance sampling, and regression analysis are reviewed for their application to the verification and validation of CAA models. Control charts are particularly useful in determining if a system is statistically within control and predictable. Innovative validation methods and techniques in the modeling community are reviewed for extension to the CAA complex simulation model. These include the traditional methodology for establishing the concept of model credibility and traditional techniques of computer code verification. Finally,

several specific large systems such as the Strategic Defense Initiative, the TAC Thunder simulation model, the Marshall Space Flight Center, and the Plant Hatch nuclear reactor are examined with respect to their verification and validation. It is possible to draw inferences from these large systems that are directly applicable to the verification/validation of complex military models developed by CAA.

B. Background.

This research follows a previous effort which resulted in a document "The Verification and Validation of Simulation Models, A Methodology," by Jerry Banks, CPT Daniel Gerstein, and CPT Sean Searles, dated September 1986. This prior research proposed a methodology that required continuous verification and validation throughout the modeling process. The resources required by this proposed verification and validation methodology were determined to be greater than that which CAA could allocate at the time to a model development effort.

In October 1987, the Director of CAA, Mr. E. B. Vandiver, conducted a site visit to Georgia Tech. In addition to the limited resources available for continuous verification and validation, the Director expressed interest in developing new and revised methods for verification and validation at his agency. The Director's interest led to a concept paper to CAA in late October 1987. Work commenced on the project in January 1988 with Jerry Banks acting as Principal Investigator. The technical contact at CAA for this research was Mr. Gerry Cooper (ARPO). Lately, Mr. Howard Whitley (MVO) has assumed much of this responsibility.

Georgia Tech has a number of Army officers enrolled in graduate programs in Operations Research, who have an interest in real Army projects. Four of these officers have assisted the Principal Investigator with various portions of the research effort.

II. Definition of Terms

This section defines the terms verification, validation and credibility as they are used in this project. Additional related terms extracted from (Banks, Gerstein, and Searles, 1986) are provided in Appendix A.

Verification: Verification refers to the comparison made between a conceptual model and the computer model employed to implement the conception. Verification techniques examine the actual computer code to ensure that it models the correct conceptual model for a simulation. Verification techniques ensure that the computer code does what it is really intended to do.

Validation: Validation refers to the techniques employed to ensure that a simulation model represents true system behavior with sufficient accuracy to allow the simulation model to be used as a substitute for the actual system. A properly validated simulation model can be used with a high degree of confidence in a series of experiments initiated to draw conclusions or answer questions about the actual system.

Credibility: A simulation model is considered to be credible if its output is reasonable and believable. A simulation model which accurately portrays the actual system under all reasonable conditions is considered to be highly credible.

Reference:

Banks, Gerstein, and Searles, (1986), "The Verification and Validation of Simulation Models, A Methodology," Georgia Institute of Technology, Atlanta, Georgia.

III. Conduct of Research

The research was conducted to develop new methodologies for the testing, understanding, and validation of complex simulation models used by CAA. The research commenced in January 1988. Initially, the research focused on three areas: (1) analogies drawn from other systems in general, (2) an examination of statistical methods, and (3) extensions of current verification/validation techniques.

The Principal Investigator focused on developing the analogies and overseeing the research conducted by CPT James Boatner and CPT Michael Casas. CPT Boatner focused on developing the extensions from current verification/validation methods. CPT Casas devoted his efforts to applying the statistical methods to the military model ATCAL. Research focused heavily on a literature review and establishing telephone contacts with a number of recognized experts in the verification/validation field. The first several months harvested a vast amount of material from the site visits and telephone contacts. This information was analyzed and synthesized into the initial draft document presented to CAA in June 1988. CPT Casas and CPT Boatner were graduated and left the research effort at this time.

Based upon the comments received from CAA on the initial draft document, a fourth area of study, inferences from specific large systems, was added to the research effort. The purpose of this research was to examine large systems to determine what verification/validation procedures, management techniques, engineering concepts, and feedback mechanisms enabled the large systems to operate well on a daily basis.

CPT John Scott commenced work on this area in October 1988. Additionally, four analogies were added to the analogies section in October 1988. CPT Donald Dawson applied regression analysis and time-series analysis to the military model TRANSMO in October and November 1988. The initial draft document was revised in January and February 1989.

IV. Analogies

A. Purpose

The purpose of this section is to examine several analogous systems which involve verification and validation and consider how these systems have application to the large-scale military simulation models commonly developed and used by CAA. After all of the analogies are presented, we draw inferences from them.

B. Introduction

Verification and validation (V/V) of analogous systems are presented. These systems include banking, physical examination, medicine and drugs, restaurants, promotion and tenure, business, religion, election for political office, building construction, medical diagnosis, Ph.D. program, food and nutrition, airline industry, personal transportation, fire fighting, and marketing. We briefly explain each analagous system, then we describe how verification and validation are practiced with implications for complex military simulations.

Some of the ideas from the analogies have already found implementation in other areas of this research. Additional analogous systems may be constructed if necessary.

Following the presentation of the analogies is a set of inferences which can be used to guide CAA in developing and implementing new and improved verification and validation procedures.

C. The Analogies

1. AREA: BANKING

Discussion:

Loans are made on the basis of creditworthiness. The lending institution must determine whether an individual is a valid credit risk. At the next level, banks are insured by the FDIC; savings and loan institutions by the FSLIC. These insuring agencies must verify that these institutions meet established standards for insurability.

a. Validating a loan prospect:

The objective of the lending agency is the return on investment. Simply stated, the lenders borrow at X% and loan at Y% where Y > X. This is an oversimplified view of the process since many firms that lend money are actually processing loans which are really made by a third party. This last statement is generally true for long term equity loans, such as for housing. Short term loans, also called consumer loans or installment loans, are usually made directly by the lending agency. However, the "paper" can be "sold" to another firm even in these instances. In any case, there must be a validation that the recipient of the funds is worthy of the risk. Factors for consideration include past history, current capability, and future capability. Past history includes prior loan repayment. Borrowers should show that they have been non-delinquent in meeting prior loan obligations. That is, the borrowers have been valid credit risks in the past. Also of concern when major loans are placed are sources of funds for the last several years. Loan processors request copies of IRS Form 1040 which shows past income. The past is evidently an indicator of the future. Confirmations of past earnings

also attained with steady employment resulting in a salary for the last several years. Borrowers that have been students and/or independently employed, or those with independent wealth, do not fit the mold and must go to extraordinary lengths to have their past history validated. Prior students may need transcripts from their schools. Independently wealthy persons will need confirmation showing the current status of their accounts.

Regardless of the risk score achieved, a loan officer can overide the quantitative information. An example of a good loan risk rated as poor is a new college graduate with an accounting degree who wants to buy an automobile. This person may not have accepted a job, has not been employed for the past year, has moved five times in four years, and so on. On the positive side, the parents of the soon graduated student have been doing business with the bank for 15 years. This individual would receive a poor rating from any of the rating services, but is probably an excellent risk and a good potential repeat customer for loans.

Analogy:

The model's past history of validity is an indicator of its current validity. It would be unusual that an invalid past would suddenly revert to a valid present and future.

b. Current capabilty relates to outstanding debts and income:

Credit card balances must not be in arrears. The borrower must currently be employed, or have a source of income. Lenders determine the borrower's ability to pay based on current income and expenses.

Analogy:

Current validity is an indicator of future validity. It is expected that the model will continue to be valid in the future, if it is valid

today. Note that once a loan is closed, there are no further checks on the borrower's ability to pay, i.e., models are not revalidated. However, if a borrower fails to meet future obligations, the loan may be foreclosed and the collateral lost. In the modeling world the analogy for foreclosure would be to discard a model that cannot be adequately validated. Continuing to use an unreliable model could damage the credibility of the modeling agency.

c. Consider installment loans:

One source of information is a data base (Credit Bureau, Equifax, and TRW are three commercial data bases). For a fee, a bona fide subscriber can determine the credit risk of an individual. Individuals in the data base are the many millions that use credit. For a particular type of credit, e.g., department store, the individual will have a number; 1 if there is a record of good repayment, a higher number if payment is less prompt. The top prospect will have all 1's.

Although scoring models have been developed for evaluating loan prospects, seasoned loan officers consider them as helpful, but not sufficient for decision making. These loan officers maker their decisions on an intuitive reaction to the loan proposal they are examining.

One variation in this procedure is offered by TRW. Their product is a credit risk scoring model, called the Gold Report, which gives the probability of bankruptcy in the next 12 months. A high derived credit risk score will alert the loan officer that a problem may exist.

Analogy:

It may not be possible to build a scoring model that can be used as the sole determinant in making decisions about a model's credibility.

Scoring models may be used to give direction in validating and verifying a model.

d. Anticipate some failures:

No matter how good the credit check is, there will be a certain percentage of loans and lending institutions that will fail.

Analogy:

Just like loans that fail, we may expect models to fail (i.e., they are invalid). The loan failure rate is anticipated. It is unrealistic to expect all models to be valid.

e. Validating a lending agency:

Lending agencies, such as banks and federal savings and loan institutions are required to meet certain obligations to maintain their good standing or validity. Those that fail to meet these criteria may lose their charters or be forcibly sold to another institution. Of concern to the federal regulators are cash reserves, loan portfolio, record keeping, accuracy of accounts, and other considerations. Some of these criteria are in terms of acid test ratios. Minimum values are set and institutions below the minimum values awaken the investigative arm of the insuring agency.

Analogy:

Acid ratios are akin to benchmarks. These ratios are standard values. Perhaps, a range of ratios for parameters can be developed for various types of models. An unscheduled model review could be initiated by an acid test ratio out of bounds.

f. Some of the criteria require deeper investigation:

Auditors certify that accounts are accurate. When suspicion is raised about accuracy, a "strike force" may pay a surprise visit to a

lending institution. Examinations are made by governmental agents, rather than by auditors.

Some of the criteria require more subjectivity. There must be assurance that proper record keeping procedures are being followed. There are standards to be followed and lending institutions are examined to determine if they are following standards.

Analogy:

Audits are similar to peer reviews, with several major differences. First, auditors are external agents, not related in any way to the lending institution. Second, auditors are looking for specific traits and records, not searching with lack of direction. For complex simulation models, reviewers could be external to the agency to assure unbiased results. Thus, CAA could review TRAC models, OCS could review CAA models, Army agencies could review Navy models and so on.

Reviewers could be directed to search in a specified manner. This could be in the form of an expert system. Thus, if the answer to A is "yes," go to B. If the answer to B is "no," go to D, and so forth. Thorough examination of past peer reviews could give direction in how this search should proceed. Additionally, interviews with former peer reviewers and modelers would expand on the search procedure.

2. AREA: PHYSICAL EXAMINATION

Discussion:

Many individuals have an annual physical examination which may consist of chest x-rays, blood chemistry, urinalysis, electrocardiogram, sigmoidscope, intensive review of the body by the physician, measurements (height, weight, body fat) and a question and answer session. This procedure has its advocates and its critics. Critics state that the exams fail to detect any significant illness because no symptoms are present. Advocates state that baseline information is important and signals can be observed that can lead to corrective methods to avoid later illness. Advocates also speak of certain illnesses that can be detected, when these illnesses are symptomless, except in advanced stages. Critics state that resources are wasted by the examinations; advocates have the opinion that preventive methods are well worth the cost.

We will examine several of the tests to see what analogies might be found. Consider the human body as the model and the various tests are attempts to ascertain validity.

a. Blood chemistry:

Commonly, one vial of blood is drawn and 18 tests are performed. Normal ranges have been developed. Warnings are sounded depending on distance from the endpoints of the normal range. The key word is "normal." What may be normal may not be healthy. The normal range may be the mean plus or minus two standard deviations. Also, values out of the range may be healthy, i.e., HDL less than 150 (normal low range value) is to be applauded.

Analogy:

The analogy in this instance is the benchmark analysis developed by McQuie. The appropriate use of the results of the benchmark analysis must be developed. Perhaps, more analysis is required before application is made of McQuie's study.

b. X-rays:

Quite different from the blood chemistry, the goal of the x-ray is the absence of spots or shadows. The appearance of a spot is rather ominous. The appearance of a shadow may require more examination.

Analogy:

The analogy here is an external glimpse of the model, peering into its innards in search of telltale signals. Display of registers, value of internal variables, and so on during a simulation may be a means of implementation. An analogy for the spots and shadows showing up on the x-ray is a fractional factorial, i.e., factors may be significant or not significant.

c. Urinalysis:

This is a very simple test which requires only the addition of a chemical to the urine. If the urine changes color, a serious problem is possible, albumen may be present. In some cases, a retest is requested to certify the original test, or to see if the problem is resolved over time.

Analogy:

The analogy here is to take byproducts of the model, and test them to be certain that inappropriate data (foreign substances) are not present. However, the test must be very simple. The analogy of a retest

is that one measure may not be sufficient. Time series data may need to be analyzed.

d. Sigmoidscope:

The test is invasive and looks into the body (model) for tumors, nodules, polyps, fissures, and lesions. Findings may be benign or malignant. Benign substances may be removed, as they may become malignant in the future, or cause discomfort. Malignancies lead to various forms of treatment. If malignancies are discovered in an advanced state, rather than an early state, the results can be fatal. The earlier the discovery, the more likely it is that a patient will survive.

Analogy:

The analogy here is an internal view of the model, but only in a specific area, e.g., the attrition algorithm. When the view is taken, there are specific warnings and alarms that can be sounded. Some alarms may be for warning purposes only, i.e., they may or may not cause serious consequences later. Some alarms are causes for immediate, sometimes drastic action, to save a model. A determination must be made concerning the tests to be made and the reason for sounding an alarm.

e. Question and answer session:

The physician asks numerous questions. If the wrong answer occurs, the physician probes further. For example, if the patient says, "I'm always tired," the physician asks how much sleep the patient is getting. If the patient is getting ample sleep, the physician tries to determine if the problem is physical or mental. The physician has vast experience in determining the appropriate questions.

Analogy:

The analogy for validating complex simulation models is the development of a standard set of questions whose answers should be in the affirmative or within a range ($a \leq x \leq b$). Negative answers or observations out of range are signals to probe further to validate the model.

3. AREA: MEDICINE AND DRUGS

Discussion:

Prior to the release of drugs in the United States, a lengthy test period is required by the Food and Drug Administration. The tests proceed from mice in the laboratory, perhaps to primates, and then to humans (prisoners, then volunteers). This procedure is very lengthy as there are many subjects and they are tracked over time. By being so meticulous, drugs that are invalid for the purpose intended may be eliminated or modified. Unintended uses may be discovered. During the testing process, side effects are noted. Some side effects are classified as major and some as minor, according to frequency of occurrence.

Complaints have been raised that the testing process is too lengthy. The process is much shorter in Europe. Those concerned with this issue argue that valid drugs are being withheld from a population that could benefit from the release of those drugs.

a. Tabulated Information:

The Physicians Desk Record (PDR) includes most all prescription and over-the-counter medicines that have been approved by the FDA and are currently available. The PDR describes the drug's purpose, dosage, major and minor side effects.

Analogy:

Indexes of combat simulation models and games have been prepared over the last 10 years. Unlike the PDR, the "side effects" of the models are not described. Side effects in the context of combat simulation would be of the following nature," will not play helicopters flying in

adverse weather conditions" or "resupply after second day of battle questioned." On a more positive note, the model's strength could also be described. Increase the amount of information provided in the catalogues of combat simulation models and games to indicate concerns related to verification and validation (how models were verified and validated, when completed, by whom) and specific statements concerning the purpose of the model.

b. Test Population:

Many subjects are used in drug testing.

Analogy:

Replication is difficult in complex simulation modeling. Push for increased replication. This may require small scale models (a mouse is a small scale of a human). A small scale model is a cut down version of a big model with either fewer modules or less complete modules. Alternately, a model with 100 weapon types might be played with 20 weapon types.

c. Error:

There is great concern of a type II error (accepting an invalid drug) by the FDA.

Analogy:

Similarly, CAA must carefully examine model subroutines, modules, and logic for type I errors, type II errors and type III errors (probability of solving the wrong model).

d. Time:

The FDA is not usually swayed by time pressure, i.e., protests that lives are being lost because of the failure to approve a drug do not decrease the testing requirements.

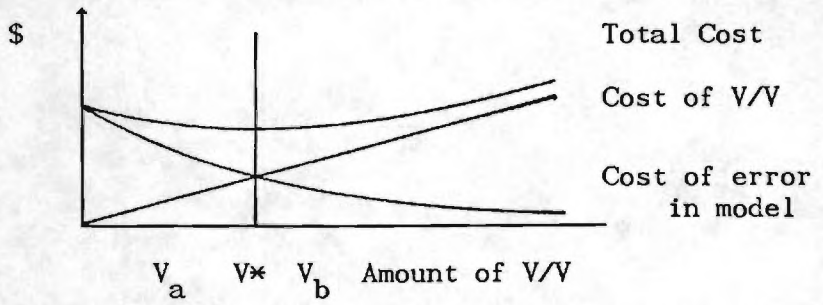
Analogy:

Development of combat simulation models is often subject to time constraints, i.e., the decision date is X months from the current time. The decision will be made with or without the combat simulation model. If completion of the model has been promised by the decision date, it is important that this obligation be met.

Insist on longer lead times in modeling. This may be accomplished by accepting tasks with longer lead times. Longer lead times will allow for more verification and validation.

e. Cost:

From an economic standpoint, consider the relationship shown below:



As the amount of validation increases its cost rises, but the cost of errors decreases. The total cost is the sum of these two costs. The minimum total cost occurs at some value V*. Perhaps the FDA is operating at V_b, because of the great concern for type II error.

Analogy:

Perhaps combat modeling agencies are operating at V_a. Perhaps a move toward V* is warranted.

4. AREA: RESTAURANTS:

Discussion:

Restaurants are validated by the public's taste and the perception of value. There are several aspects to the public's taste, and taste differs. A five year old child would probably prefer a hamburger and French fries from McDonalds than dinner at the Ritz. Limiting our concern to restaurants that serve fine foods provides a homogenous set for analogy development. Restaurants are reviewed in newspapers and slick magazines, and the health department determines if sanitation laws are followed. Recommendations of friends provide confirmation that a restaurant may be valid. The restaurant's menu is a statement of what is intended, i.e., the restaurant's objectives.

a. Restaurant reviews:

The critics go further than determining that the meal does or does not taste good. They are concerned with ingredients, appearance, texture, adherence to standard recipes, innovation, freshness, and other criterion. They are also concerned with restaurant decor, promptness of service, expanse of wine list, and other criterion. Reviewing seems to be an art, depending on the experience of the reviewer. The reviewer generally makes several visits rather than a single visit and may sample several dishes, or try the same dish a second time in a test for consistency. Virtually everyone eats food, and we have our likes and dislikes, but we are not all food critics who will be respected for our opinions.

Analogy:

The analogy in combat simulation is an expert modeler who knows just what to look for when evaluating a model. This could be called a peer review by a committee of one. Peer reviews seem to be an art, just like restaurant reviews. The only difference is that peer reviewers seem to be on a hunting mission rather than directed. A peer review with structure is suggested. Detailed instructions will be provided for conducting a peer review, but these instructions are modifiable by the leader of the review team.

b. Health department:

Sanitation laws are drafted by the local government and the health department determines if the rules are being followed. There are various categories and points are given in each category. A score of 100 is perfect. A score less than the minimum acceptable, say 80, will result in the restaurant's closure. A score between, say 80 and 90, may result in probation. A return visit is scheduled. If the restaurant fails to score over 90 on the return visit, closure may result.

Analogy:

The analogy in simulation of combat is a scoring model which gives (verification and validation) points in certain categories. To be declared credible, a model must score above some minimal value. A procedure of this nature was suggested by Gass. Others have offered similar scoring models, but combat modeling agencies have been reluctant to implement such models. One problem is the arbitrariness of scoring models. Finding an implementable scoring model is the task to be overcome.

All too often, little verification and validation is performed on complex simulation models. That which is accomplished is usually in the form of a peer review which lacks structure and thoroughness, and is too dependent on the reviewer or reviewers. These peer reviews are primarily performed after a model is completed. Peer reviewers are usually limited in number, limited in time, and cannot understand every algorithm in a big model. Because of these constraints peer reviews must be superficial, and, perhaps, opinionated rather than factual.

c. Specifications:

The restaurant's menu is a statement of what outputs are promised. A restaurant that cannot deliver its menu items is impotent and a critic is not required to determine if a menu can be followed or not.

Analogy:

In combat models, it is necessary to return to the original specifications (detailed statement of objectives) to determine if a model is delivering as planned, or as modified.

d. Recommendations:

Recommmendations by friends (word-of-mouth) spread rapidly. A chic, new restaurant with outstanding food at reasonable prices soon has waiting lines for dinner.

Analogy:

If the users like the model, they tell other potential users of the model's worth. The analogy here is to create an electronic bulletin board with which model users could communicate their experiences, favorable and unfavorable, concerning a combat simulation model.

5. AREA: PROMOTION AND TENURE

Discussion:

The promotion and tenure of a professor on the faculty of a major university is determined over time by demonstrated performance and future potential. A number of different factors affect a professor's promotion and eventual tenure.

a. Multiple Decision Criteria:

The promotion and tenure system in a major university is based on teaching, research sponsorship, scholarly productivity, service, and outside reviews.

Analogy:

Use many characteristics before making a decision about model credibility. Require a minimum response for each type of information, i.e., minimum requirements as a teacher, etc. Thus, exceeding the requirements in one area cannot eliminate meeting the minimum requirements in another area.

b. Multiple inputs for determining promotion/tenure:

There is input at the school or department level, then the dean's level, then the academic vice president's level, the president's level and finally at the board of regents level. Although the names of these academic bodies vary from institution to institution, roughly the same number of levels must be negotiated. In addition, some of the levels may have several sources of input. For example, at the school or department level there may be peer input to a school or department committee that provides input to the director or chairperson who formulates a decision.

Analogy:

Use many information sources. The use of multiple information sources reduces the chance of a type II error, accepting a false hypothesis.

c. Reject at lowest level:

At some levels, a negative conclusion can stop the promotion or tenure proceedings.

Analogy:

Allow decisions to be made in sequence so that they do not have to be made by the last person in the chain. The last person is the least likely to have a complete grasp of technical details and issues as this person is responsible for many types of decisions and the solution of many ongoing problems. (Note that this is quite the opposite of the peer review process in a combat modeling agency which calls upon the top person to make a final decision.

6. AREA: BUSINESS

Discussion:

Businesses are verified and validated by reputation, customer satisfaction, profitability, and continuity of service among others.

a. Credibility:

The credibility of a business is a function of many indicators. Some of these are acid ratios. For example, the price to earnings (PE) ratio of a stock is reported along with the price of the stock and its high-low price. Investors look for low PE ratios as opportunities.

Analogy:

Similarly, complex simulation models could be subjected to acid test ratios. One step in this direction is the benchmarks analysis conducted by McQuie.

b. Uniform performance:

Corporation offices expect individual businesses to perform similarly in various locations. Test marketing of new products is conducted to determine what will occur when the product is introduced to a larger market.

Analogy:

In V/V of a military model we should expect to achieve similar outputs when the inputs are varied slightly. Also, a central authority must take control of the model and allow authorized variations only. Replications of the model must be nearly uniform.

c. Bookkeeping:

Proper bookkeeping is necessary for the success of a business.

Analogy:

The analogy in V/V of a complex simulation model is documentation. As there are proper bookkeeping and accounting methods, there should be proper documentation methods. As there are auditors who insure that proper bookkeeping and accounting methods have been followed, there should be those who insure that proper documentation standards have been followed. The auditors who certify the accuracy of the accounts in a business are from an external firm, and the same modus operandi should hold for those who ensure that documentation standards are being followed. That is, an external agency should be responsible for determining that documentation of large scale simulation models is proceding during the software life cycle.

Rather than review the entire documentation effort on a continuous basis, it would be sufficient to sample the documentation effort periodically. If sampling indicates that the error rate is acceptable, there is no need to go deeper into the documentation. However, if there is an indication that the errors are excessive, rectification of the errors would be required. Rectification is accomplished by complete review of the documentation.

7. AREA: RELIGION

Discussion:

There are numerous religions. People grow up with certain cultural and religious practices or select the religion that most closely meets their personal needs and beliefs.

a. Credibility:

Religious beliefs are based on a "leap of faith." This means an individual accepts the religion (model) without questioning.

Analogy:

It is important that a model has credibility so that its use is unquestioned. The GAO (1987) report defines a simulation's credibility as the level of confidence in its results. "To say that simulation results are credible implies evidence that the correspondence between the real world and the simulation is reasonably satisfactory for the intended use. Credibility is not an absolute condition but measured on a continuum."

b. Many right answers:

There are many religions (many models) and most cannot be proved right (validated) or wrong (invalidated). There is not just one valid religion/model, but many.

Analogy:

More than one model may solve the same set of objectives.

c. Objective:

We must think of religion (the model) from the standpoint of the objective. The objective is to calm the practitioner's fears about life

and beyond, and to bring order out of chaos, i.e., provide a systematic and recognizable base.

Analogy:

Models must be evaluated with respect to their previously defined objectives. More than one model can solve the same set of objectives.

Reference:

"DoD Simulations: Improved Assessment Procedures Would Increase the Credibility of Results," (1987), GAO, Washington, D.C.

8. AREA: ELECTIONS FOR POLITICAL OFFICE

Discussion:

There are many ways that campaigns proceed and the electorate makes its decisions. Consider these factors: past record, advertising, media posture, comparison to individual's political preference, personal prejudice, constitutional requirements such as age and citizenship, comparison to other candidates, and public opinion polls. Several of these subtopics can have application to simulation modeling.

a. Public opinion polls:

Consider exit polls that are conducted in which voters are asked how they voted. These polls are more accurate than polls of voter preference since many people do not vote, even though they say they intend to vote, and voters change their minds.

Analogy:

The analogy for verification and validation of large scale simulations is to sample pieces of the code rather than the entire code. This may result in as much information as examining the entire code. It may be better than examining the entire code as modelers become very bored with a line-by-line search of a vast amount of code. (Quality controllers have shown over and over that sampling is better than 100% inspection, i.e., more errors are discovered through sampling than with 100% inspection.)

Stratified random sampling may be used. The stratification can be on the basis of importance, with a limitation on the amount of resources to be utilized. Alternately, random sampling may be used with the

understanding that this procedure assigns equal weight to all portions of the code.

b. Comparison to Other Candidates (Models):

Consider the case in which a new model is replacing an old model. (A fresh candidate goes against an existing office holder with the fresh candidate offering new alteratives.)

Analogy:

Output values over a range of input values are required and a pairwise comparison is performed. (This is true for evaluating competing political candidates and can be applied to large simulation models.) If the new model performs as well as the old, and provides new features, we might be inclined to accept the new model. If the new model performs as well as the old model, but there are no new features, keep the old model. The problem is ascertaining that the old model was valid. If the old model has been in use for many years, and decisions have been made using its output, the old model is valid. It is not feasible to "go back in time" and validate every model that is currently in production, but not properly validated prior to its implementation. Rather, cut losses and initiate appropriate validation and verification procedures now!

c. Personal Prejudice:

In the election process, the voters can either select a candidate, or reject a candidate by voting for another candidate.

Analogy:

In the modeling world we cannot always reject a candidate so easily. We may be forced to validate and verify an existing model if there are no substitutes for it.

d. Background:

In the election process, the media provides information about the candidates background, all the way to the candidates childhood. In this way we can see how the candidate arrived at the posture taken.

Analogy:

Similarly, in the model world, we can trace a model back to its beginnings, looking at the physical principles, the calculus, of its conception.

9. AREA: BUILDING CONSTRUCTION

Discussion:

An inspection occurs for each major subsystem. Consider a residential home. The foundation, plumbing, electricity, and completed dwelling are inspected at various times. The lender, or guarantor of the loan, also performs an inspection. The prospective homeowner may also commission inspections to be performed. Local standards are the criterion for some of the inspections. The guarantor of the loan, such as the VA, has another set of standards. The homeowner has another set of standards, i.e., an agreement with the contractor. With respect to V/V, we have the following:

a. Standards:

Standards are written documents which all builders must follow. At least, the minimum standards must be met.

Analogy:

Similarly, thorough, unambiguous standards can be developed for verification and validation of large scale military simulations.

b. Subsystem validity:

Each subsystem is subjected to standards as that subsystem is completed during the construction phase. If the subsystem fails to meet the standards, it must undergo revision as needed.

Analogy:

Similarly, each module can be subjected to V/V in a complex military simulation.

c. Multiple standards:

There are different standards for different parts of the nation.

Analogy:

Similarly, there can be different standards for different types of simulation models, i.e., the purpose and objectives of the model dictate the validation and verification procedures required.

d. Who sets the standard:

Standards are developed by private groups and adopted by municipalities.

Analogy:

Similarly, standards for complex military simulations can be developed by joint task forces from all services and adopted by the individual services.

10. AREA: MEDICAL DIAGNOSIS:

Discussion:

Physicians must execute a thorough and comprehensive set of procedures when making a medical diagnosis. They rely upon their training, their experience, intuition, questioning of the patient, examination of the patient, review of medicial literature, and medical testing to reach their conclusion.

a. Past history and routine analysis:

The common way to diagnose a complaint is for the physician to recall similar symptoms and correct diagnoses. He utilizes routine tests and procedures, initially, to aid in diagnosing the illness.

Analogy:

We can develop primary and secondary analytic tests for V/V. These are similar to the physician's analysis by asking the proper questions, using a stethoscope, looking at the patient's throat, and taking blood pressure--all tools that have been in use for many years. By asking the right question of models, and by examining a few key outputs we should be able to diagnose a very large proportion of the problems that exist. Just as the physician is trained in these analytic procedures, sees many patients and makes many diagnoses to develop skill, we should develop the same kind of skill among the modeling community.

b. Advanced methods:

Increasingly, physicians are relying on lab tests, invasive scopes, x-rays, cat scans, and combinations of these methods to assist in diagnosis.

Analogy:

When the primary analytic tools are insufficient to the task, we need more advanced tools to diagnose models, i.e., perform V/V. Invasive scopes may be called for in which a model is opened and the insides examined. Traces may be required in some instances where it is necessary to follow a fluid through one subsytem of a model. Usually, much preparation is required for a trace, i.e., isolating that subsytem (module), allowing no extraneous data in the system which is analogous to restricting the diet of a human prior to any gastrointestinal examination).

c. Expert systems:

Expert systems have been developed to aid in the diagnosis. These applications include and exclude certain diseases as more symptoms are provided, until the physician is left with only a very few choices.

Analogy:

Work should commence on using expert systems to diagnose complex military simulation models. This expert system would be rule based and guide modelers in the proper conduct of V/V.

11. AREA: PH.D. PROGRAM

Discussion:

There are a number of hurdles which must be passed. The qualifying examination certifies that an individual has the background to begin Ph.D. level study. The comprehensive written exam certifies that a person is ready for Ph.D. level research. The oral exam, usually taken shortly after the comprehensive written exam, is another hurdle prior to the beginning of dissertation research. Questions of a very general but basic nature are asked. The next nettlesome burden is the dissertation topic proposal. Finally, at the conclusion of the dissertation is its defense. In every case, the examination and subsequent decisionmaking is made by the appropriate committee. Failure can come at any point. The hurdles are numerous and of several different forms. The tests are spaced over time. If a test is failed, another try may or may not be authorized, but only after the proper remedy has been taken, for example, take two more courses in the weak area. Retests are only allowed when there is optimism that the student will succeed on the next attempt. The Type II error (accept a false hypothesis) is small. The Type I error (reject a true hypothesis) may be large.

a. Multiple testing:

Analogy:

Models must be subjected to many and varying types of tests. These tests are conducted over time at various stages in the development of the model.

b. Verify each level:

Analogy:

Another level of model development does not begin until the test results are satisfactory at the previous level.

c. Realistic Determination of Capability:

Analogy:

Accept reality. If a model is a failure, and too expensive to correct, there is no need to argue for its continued development.

12. AREA: FOOD AND NUTRITION

Discussion:

We verify and validate food by taste, appearance, nutritional reports, satisfaction derived from consumption, what other people say and do, critic's advice, the mood we are in, minimum nutritional requirements, and by other methods. Some of these methods may have implications for V/V of large scale simulation models.

a. Reviews:

Restaurant reviews describe the experience of experts. The reviewer's tell us more about the foods than we could even think to ask. They try to make the review more than a determination of whether or not the food tastes good. How does an individual become a good reviewer? Some of the best are also very good cooks.

Analogy:

The analogy for large scale military simulation models is in the use of an outside or independent expert who has no vested interest in the model. The best critics may be modelers or former modelers. The critics review numerous models, not just one model or one type of model. The critics must look below the obvious superficial aspects of the model, i.e., face validation is insufficient.

b. Minimum Requirements:

Nutritional requirements establish "minimum daily requirements" of vitamins and minerals. These are established by committees and change as more is learned about bodily needs.

Analogy:

In V/V of large scale military simulations we can do the same, i.e., minimum required V/V standards (tests, reviews, etc.) can be established. These standards should apply to a class of models. They cannot be generic, so there may have to be a different set of standards for each model type, i.e., interactive models, training models, analytic combat models, etc.

c. Fortification:

Nutritional requirements dictate that foods be fortified to provide essential nutrients. For example, some breakfast cereals are fortified with essential vitamins and minerals.

Analogy:

We may determine that some models are anemic and need to be fortified (develop faster algorithms, use more efficient algorithms, use double precision, etc.)

13. AREA: AIRLINE INDUSTRY

Discussion:

The major airlines in this country are continually validated by the service they provide. While there are countless variables over which they have no control, one area that can be influenced is the performance of the pilot/aircraft system.

a. The pilot:

The airlines' major source of pilots is the military. By the time an individual seeks employment with a carrier, he or she has served a minimum of seven years on active duty (including flight school) and accumulated 1500-3000 hours of "stick" time depending on the type of aircraft flown. Once hired by the airline, the new pilot spends about three months in training to become a flight engineer on a B-727. This training includes a month of ground school, six weeks of simulator training and two weeks of commercial flights with an instructor present. He or she is then assigned to a crew and placed on probation for one year, subject to dismissal for safety violations or unsatisfactory progress. Once his probation is completed, the pilot is required to have a flight physical and pass a "check ride" (in both a simulator and aircraft) annually. This pattern of revalidation continues as the pilot progresses from flight engineer to co-pilot and upgrades to larger aircraft. Once promoted to the position of captain, the pilot is still required to have a physical each year, but must take a check ride semi-annually.

Analogies:

1) A model should be revalidated at projected intervals. Moreover, the revalidation effort should be conducted at different levels. The entire model, as well as separate modules, need to be reevaluated. Based on new data from field tests, system performance and/or expert opinion, the model or one or more of its components may be inadequate or inaccurate.

2) Models should be designed with diagnostic features. If the performance of a module or subprogram can be monitored separately, revalidation will be quicker and more comprehensive.

b. The aircraft:

Once an airplane is in service with a carrier, it is subjected to three levels of inspections for serviceability. After each 3000 hours of flight time, the aircraft is virtually torn apart. The wings and fuselage are x-rayed for cracks; the engines are removed and undergo a radiation (isotope) inspection for weak components; and the flaps/pylons are removed and overhauled. The 3000 hour service takes a minimum of three weeks and occurs more frequently for the larger aircraft. (On the average a 727 flies 10-14 hours a day, while a 747 averages 12-18 hours a day.)

Every 300 hours the aircraft is taken off the flight line and inspected by a maintenance team. This inspection normally takes about eight hours to complete and includes a complete lubrication of all moving parts.

Finally, every 24 hours, the aircraft is inspected by a mechanic. This occurs on the flight line and lasts around an hour. The focus of this inspection is to look for obvious deficiencies on the aircraft and

insure all electrical systems are within tolerance. Additionally, the flight engineer is required to inspect the aircraft prior to each flight.

Analogies:

1) Analysts should also be revalidated periodically. Furthermore, diagnostic tests of the analyst could be included in the model's design. While such an approach would mainly address the objective nature of the model, it could also reduce errors in the interpretation of output.

2) Analysts should be exposed to different size/scale simulations. Proficiency with only one type of model could encourage a myopic view of a simulation and limit subjective interpretation of the output.

14. AREA: PERSONAL TRANSPORTATION

Discussion:

Consider the businessperson who moves to a new city and must find a reliable way to get to work. Through time, the businessperson will move from the novice, who cannot find the corner grocery store, to the expert who knows the best ways to get around town.

a. Transportation selection:

The new driver searches for a mode of transportation that is quick, safe, and reliable. Perhaps it is a train, subway, carpool, private automobile, bus, or walking. The relative trade off of costs associated with each mode of transportation (dollars, time, convenience, etc.) is considered.

Analogy:

Study the model. Determine an initial verification/ validation plan that best optimizes dollars, time, and standards required.

b. Vehicle selection:

Let us assume the businessperson decides to ride each day to work by personal automobile. A car must be selected that is reliable, safe, not too expensive, easily maintained, and fuel efficient.

Analogy:

Select V/V methods that have proven themselves to be reliable, time-efficient, manpower efficient, relatively inexpensive, and understood by the analysts.

c. Route selection:

Now that he or she has a car, the businessperson must select the best routes to drive to get to work. Travel speed, traffic congestion,

travel time, and road conditions are considered in any decision. Business associates, friends, and neighbors provide information on which routes are best to take.

Analogy:

The V/V team should consult with model users and other V/V experts when determining V/V methodology. Additionally, this consultation helps them to get familiar with model characteristics, quirks, and shortcomings.

d. Reference map:

Once the businessperson has selected a route, a map is placed in the car for ready reference should he or she get lost or be forced to detour.

Analogy:

Create a base document (road map) that outlines the V/V process and methodology as it will apply to the particular model. This document should contain time tables for key events in the V/V process and suspense times. It should outline the V/V process in sufficient detail to erase any ambiguity for the analyst. It should be referred to frequently and amended as necessary.

e. Route execution:

Through trial and error, the businessperson drives the selected routes and learns first-hand their strengths and weaknesses. Decisions are based on which routes are most suitable for varying traffic conditions and different times of the day. Alternte routes are identified that can be used during road closures and traffic backups on the primary thorough fares.

Analogy:

The V/V team must execute the model as time and resources permit. Examine outputs and how they differ based on different inputs and changes in model parameters. Examine individual modules. Perhaps some are redundant. Perhaps a bad module can be rewritten or replaced by a known module of better quality.

f. Routing updates:

During the drive to work, the businessperson listens to the local radio station to get updates on traffic watches in the city. These reports assist the driver to steer clear of accidents, chokepoints, and traffic congestion.

Analogy:

The V/V team should stay abreast of what others are doing in their field. Perhaps techniques found to be successful on concurrently running projects can be applied to this model. Invite outside experts to examine the work done thus far for accuracy.

g. Route update:

The businessperson utilizes new roads as they are built to make commuting shorter, faster, and more efficient.

Analogy:

Utilize new V/V techniques as they become available. Keep the analysts up to date and educated on these new methods.

h. Emergency services:

Commuters familiarize themselves with the location of key emergency services such as police, hospital, and vehicle repair. Automobiles are properly maintained, fueled, and licensed.

Analogy:

Be prepared to call for outside help if the current V/V team runs into trouble. Be prepared to bring in additional resources or more robust techniques to solve tough problems.

i. Record keeping:

The businessperson keeps a record of car expenses for tax purposes and reimbursement.

Analogy:

Document all results and changes made in the model through the V/V process. Ensure these documented results are easily understood by new analysts added to the project. Ensure the results of tests on the model are documented sufficiently so that repeated tests bear similar results.

15. AREA: FIRE FIGHTING

Discussion:

Consider the personnel selection, training, equipment selection, and rehearsal required to prepare fire fighters to do their jobs.

a. Mission:

When a fire station is built it is designed to serve a specific purpose. The fire station must be capable of extinguishing a certain size of fire according to the type of fire. For example, a fire station in a residential area must be capable of successfully extinguishing a wood based fire in residences up to 3,000 square feet with personnel and equipment assigned. This particular fire station may not be capable of extinguishing a petrochemical blaze at a nearby gas station without special help. In accordance with its mission, the fire station is assigned certain types of equipment and a certain number of fire fighters, each with certain fire fighting skills.

Analogy:

Examine the model to be validated/ verified and tailor the team of experts who will examine it to fit the needs of the model. Ensure that adequate numbers of analysts are assigned and that they possess the skills needed to service a particular model. Ensure they have the computing facilities and analytical tools on hand commensurate with the size of the job.

b. A fire occurs:

When a fire breaks out in a family residence the occupants must react quickly if they are to save the dwelling. They should call in the

fire station to help rather than attempt to estinguish the blaze themselves.

Analogy:

Testing and analyzing the model should be left to experts.

c. Initial decisions:

When the fire fighting team arrives they must quickly assess the situation. Are all occupants of the house accounted for? If someone is trapped inside, where are they? Are there any particularly hazardous materials in the house (e.g., car with full fuel tank in the garage, fertilizers, chemicals, and paints stored in the garage)?

Analogy:

The evaluation team conducts an initial review of the model. They determine what is wrong with the model and what needs to be fixed first. The team must be familiar with what the model is supposed to provide in terms of output and capabilities.

d. Utilize available resources:

Before the fire fighters use the water on their trucks, they attempt to find local water mains to fight the fire.

Analogy:

Identify what resources the local model users can provide to assist in the V/V effort. They have historical knowledge of how the model works and historical data files of previous model runs. They may have personnel who can assist the V/V team with certain aspects of analysis.

e. Execute rehearsed plans:

The fire fighters act as a team. Each firefighter has a job to do and executes it quickly. Through teamwork they accomplish the task faster, more efficiently, and safely.

Analogy:

Ensure one analyst is in charge of the V/V effort. Ensure that the V/V process follows a detailed plan to maximize time and resources. Utilize proven V/V methods to analyze the model.

f. First things first:

The firefighters attempt to contain the fire first to keep it from spreading to other buildings or property. They then begin to reduce the fire until it is safe enough for firefighters to enter the building, if necessary, to put out hotspots.

Analogy:

Apply the V/V process systematically under the supervision of the team leader. Ensure that personnel are not duplicating each other's work. Ensure that each member of the team knows what is going on and is aware of the progress being made.

g. Call for help if needed:

If the fire cannot be contained with the assets on hand, then a decision must be made quickly to bring in additional resources.

Analogy:

If the model cannot be analyzed with the resources on hand then additional experts, computing facilities, software, or techniques may be required. Bring in the necessary assistance to get the job done before too many resources are expended needlessly.

h. Follow up:

Once the fire is out, experts come in to examine the dwelling to determine the cause of the blaze. The fire fighters review their efforts and determine ways to improve their responsiveness and fire fighting techniques.

Analogy:

Upon completion of the V/V effort the process should be documented. Lessons learned should be identified and disseminated to other V/V teams. Particularly promising analytical techniques should be further refined for future use.

16. AREA: MARKETING

Discussion:

Few companies can conduct their business successfully without forecasting the future demand for a product. Unfortunately, not many products or services lend themselves easily to such an activity. In the majority of markets, demand is not stable from one time period to the next; accurate forecasting is therefore a key contributor to company success. Otherwise, forecast errors can lead to excess inventory (and expensive product markdowns) or lost sales opportunities when a product is out of stock. The more volatile the demand, the more critical the forecast.

Forecasting methods range from the crude to the highly complex. All have one trait in common, they are built upon an information base. Furthermore, there are only three bases upon which to build: what people say, what people do, and what people have done. The first basis (what people say) includes the approaches of (1) surveys of buyer intentions, (2) sales-force opinions, and (3) expert opinion. Basing a forecast on what people do "simply" subjects the product to a market test in order to indicate buyer response. The third basis (what people have done) mathematically and/or statistically analyzes past records of buyer behavior; two common methods are time series and statistical demand analysis.

a. Surveys of buyer intentions:

This method has several limitations in practice. Buyers do not always freely report their intentions, and when they do formulate intent, often it is not carried out. As a result, this approach works best for

consumer durables, product purchases where the buyer must plan in advance and new products where past data are not available.

Analogy:

Given no system history, how much of the model is based on assumptions? Does any empirical data exist on which to base these assumptions? Since predicting future behavior is highly subjective, a model's margin of error may be too large to be termed valid. In other words, if the nature of the system forces the model to rely too heavily on assumptions then perhaps the system cannot be simulated.

b. Sales force opinion:

Few companies use their sales force's estimates at face value; these are biased observers. They are often unaware of "the big picture" or may underestimate demand hoping for lower sales quotas. However, assuming these biases can be corrected, certain benefits can be obtained. Sales representatives are usually close to developing trends; participation in the process tends to increase incentive to achieve; and customer input is accounted for indirectly. This approach is advantageous when the sales force is the most knowledgeable source of information.

Analogy:

V/V efforts must include the model builders; responsibility for the model should permeate this process. While the modelers will be biased, their input can be made more accurate through incentives.

c. Expert opinion:

Marketeers use at least three ways to gather information from outside experts. They meet as a committee and produce a group estimate. They supply individual estimates to a designated leader who produces a merged, single estimate. Or, iterative sets of individual estimates are

submitted until they converge to a consensus (Delphi Technique). The use of expert opinion is advantageous because forecasts can be made quickly and inexpensively. Also, different points of view are surfaced. The major disadvantages are that responsibility is dispersed (or non-existent) and "bad" estimates are given the same weight as "good" estimates.

Analogy:

As stated earlier, peer reviews should utilize impartial experts. Moreover, these reviews should be conducted sequentially until a consensus is reached. "Bad" reviews need to be reconciled prior to each subsequent iteration of the V/V process.

d. Market tests:

In cases where buyers do not plan their purchase(s) or are unpredictable in carrying out intent, a direct test of consumer behavior is desirable. Because these tests are normally conducted on a small scale, the best results usually occur for short-term estimates.

Analogy:

Whenever the system is unpredictable (as most are), the V/V process is strengthened by participation of model users. A corollary to this analogy is that the more often user observations are addressed as they occur (rather than after the process is completed), the more closely the model will simulate the system.

e. Statistical analysis:

Time Series - this approach treats past and future sales as a function of time. The underlying object being that sales levels are an expression of enduring causal relations that can be expressed quantitatively. Normally, a time series will account for a combination

of four components: trend, cycle, season and/or erratic behavior. These components may interact linearly or multiplicatively, but the goal of a time series analysis is to avoid extrapolation. The major disadvantage to this method is that no stable relationship may exist.

Regression - When demand factors are unstable (in relation to time), it is often more efficient to ascertain the relationship between sales and demand. Statistical demand (regression) analysis attempts to discover the most important factors contributing to the variation in sales. The procedure expresses sales as a dependent variable in terms of the variation in any number of independent variables (i.e., prices, income, population, promotion). The problem with regression analysis is that the demand equation's validity can be diminished by too few observations, too much correlation among the variables or violation of normal distribution assumptions.

Analogy:

If the model has past performance data, how was it initially validated? Did the V/V process include numerous observations or were results extrapolated from isolated data? The more strenuous the effort, the greater the chances for future validity.

D. Inferences from Analogies

Each analogy offered several constructs that could be employed in the validation and verification of complex military simulation models. These constructs, or inferences, are presented in this subsection.

Banking

1. The model's past history of validity is an indicator of its current validity.
2. Current validity is an indicator of future validity.
3. A certain proportion of models will be invalid.
4. A model review can be initiated by an acid test ratio out of bounds.
5. Peer reviewers should be members of external agencies.

Physical Examination

1. Normal ranges for benchmarks can be established and warnings can be sounded when an observation is out of bounds.
2. An internal glimpse of a model can be attained by displaying certain registers and values of internal variables during the simulation. These can be compared to baseline values.
3. Data at one point in time may not be sufficient. Time series data may be more important.
4. Only selected algorithms of a model should be examined with a specific interest in mind. Results may call for drastic action.
5. A standard set of questions and acceptable responses should be developed. Answers out of range indicate that further probing is required.

Medicine and Drugs

1. Push for increased replication even if small scale models of larger versions must be used.
2. Develop training materials that explain basic statistical concepts. Offer training in statistical methodology.
3. Insist on longer lead times in modeling.
4. Increase the budget for V/V.
5. Increase the amount of information available to the DoD community on the V/V of models.

Restaurants

1. Provide structure for peer reviews.
2. Search for an implementable scoring model.
3. Compare model outputs with that which was established in the original specifications as modified.
4. Create an electronic bulletin board that would allow model users to communicate their experiences.

Promotion and Tenure

1. Use many pieces of information before making a decision about model credibility. Require a minimum response for each type of information.
2. Allow decisions about model credibility to be made in sequence so that the last person in the chain does not have to be the decision maker.

Business

1. A central authority must take control of the V/V process and authorize all modifications to the model.
2. An external agency should be responsible for determining that documentation is proceding during the software life cycle.
3. The quality of documentation can be ascertained by sampling.

Religion

1. Models must be evaluated with respect to their previously defined objectives.
2. More than one model can satisfy the same set of objectives.
3. Practitioners will usually use the model with which they are most familiar.

Elections for Political Office

1. Verify using pieces of code rather than the entire code.
2. It is not feasible to validate every model in production. Rather, cut losses and initiate appropriate V/V procedures now.
3. Trace models back to their conception, looking at physical principles.

Building Construction

1. Minimum standards must be developed for large scale military simulations.
2. Each module should be subjected to V/V.
3. Different standards should be developed for different types of models.

4. Standards can be developed by joint task forces and adopted by the individual services.

Medical Diagnosis

1. Persons possessing skills in V/V should be trained through continuous assignment to a model evaluation team.
2. Work should commence on using expert systems to diagnose complex military simulation models.

Ph.D. Program

1. Models must be subjected to many and varying types of V/V efforts at numerous stages in their development.
2. The next level of model development should not begin until the current level has been verified and validated.
3. If a model is a failure, stop the development process.

Food and Nutrition

1. Use outside or independent experts that have no vested interest in the model for some portions of the V/V exercise.
2. Establish minimum standards for V/V. Different sets of standards must be developed for different models.
3. Some models may need fortification (faster algorithms, more efficient algorithms, double precision, etc.).

Airline Industry

1. Models should be revalidated at different levels of scrutiny.

2. Models with internal diagnostic capabilities would streamline the revalidation process.
3. Vehicles for analyst revalidation should exist where objective portions of a model are concerned.
4. Analysts should be proficient with numerous types of models.

Personal Transportation

1. Tailor a specific V/V plan to the model. The plan must organize personnel, resources, and time to accomplish V/V in a methodical process.
2. Use proven V/V techniques. Create a road map (base document) that outlines the V/V process with projected suspense dates and assigned responsibilities.
3. Consult with model users and experts when developing the V/V plan.
4. Run the model, as resources permit, with various changes in input data base to determine model outputs.
5. Invite an outside look to measure progress and accuracy of V/V work to date.
6. Use outside experts and resources if needed.
7. Document all results and recommendations of the V/V process to include changes made to the model.

Fire Fighting

1. Tailor the V/V team of experts to fit the requirements of the model. Ensure that necessary hardware and software are available to fit the task.

2. Identify resources and information the model users can provide.
3. Ensure overall responsibility for the V/V effort rests with one individual. Delineate takes and responsibilities clearly.
4. Apply V/V systematically and avoid duplication of effort.
5. Document all lessons learned.

Marketing

1. The model should not be based solely on assumptions; some systems cannot be simulated.
2. Model builders should be included in the V/V process, with their input reflecting responsibility for the model's validity.
3. Peer reviews should reflect an iterative consensus.
4. Model users contribute most to the model's validity when their observations are addressed concurrently with other V/V activiities.
5. Past (or current) validity of a model can be deceiving. The more thorough the validation process, the greater the probability of future validity.

V. Large Systems

A. Purpose

There are exceptionally large and complex systems in operation that demand precise specification of design, performance, and output. For example, oil refineries, nuclear reactors, the space shuttle, and an aircraft carrier are all examples of complex systems that work well on a daily basis. The oil refinery must produce a myriad of products to design specification while maintaining efficiency of production and profitability. A nuclear reactor must be designed to produce power efficiently and in an inherently safe manner. The space shuttle is an extremely complex machine utilizing the latest engineering design technologies. It must operate safely with almost no margin for fatal or irreversible error. An aircraft carrier harnesses the skills of 4000 persons to accomplish its mission. All of these systems are extremely complex, yet all function very well on a continuous basis. It should be possible to examine large systems like these, identify the verification/validation techniques used during their development, and apply the resulting inferences to military models.

B. Introduction

In terms of their simulation modeling, large systems can be classified into two broad categories. The first category is the large system for which historical data on system output and performance is readily available. The second category is the large system for which historical data on output and performance is not available. An example of the first system is a power generating nuclear plant for which a training simulator for operators will be developed. For a given initial data base of operational inputs and parameters, the simulator output can

be compared directly to actual plant output when operated under the same initial conditions. Comparison of simulated output to actual plant output can be readily made and the accuracy and credibility of the simulation model can be judged in a straightforward manner. An example of the second category of large systems is theater level war in Central Europe in the 1990's. No such war has been fought. Consequently, no historical "real war" data base on the success of combat operations with current equipment, doctrine, and training in Central Europe can be referenced. Thus, the output of a simulation model developed for a war in Central Europe is harder to quantify as being accurate and acceptable for reliable force planning. The need to verify and validate such a model does not diminish, however, and creative methods must be developed to establish their credibility. In this chapter several large systems from both categories will be examined to describe creative techniques and strategies of verification/validation. From these techniques and strategies, inferences can be drawn to improve the verification/validation of military simulation models.

C. Large System: Plant E. I. Hatch Nuclear Power Plant

1. Discussion

Plant E. I. Hatch is a power generating nuclear reactor in Baxley Georgia. Within the nuclear industry it has the reputation for having an excellent training simulator.

a. Simulator Standards

Within the nuclear industry, computer simulators have played an increasingly important role in training plant operators. Due to the wide variety of plants in operation it has become necessary for nuclear power plant simulator standards to be developed. These standards specify the minimum requirements for simulator performance and configuration necessary for effective training. The American National Standard ANSI/ANS-3.5-1985 provides these standards for the nuclear industry. Generally, this document requires that the steps taken by the plant operator to run the simulator parallel those taken to run the actual plant. Simulator response must be realistic enough that the operator would not observe a difference between the simulator control room instrumentation and the reference power plant control room. The simulator must display operating conditions on control panel displays analogous to the power plant. The simulator must have the capacity to store twenty sets of initialization conditions. It must be possible to start and stop a simulation in progress to insert various malfunctions to test operator response (man in the loop). Freeze simulation capability is required. Fast time, slow time, backtrack, and snapshot capabilities of simulator response are considered to be important. Instructor interface with the simulator is required to allow intervention into the

simulation. Hardcopy transient data during simulation is required. Various annunciators and indicators of power plant operating conditions are built into the simulator. Verification/validation requires that the simulator be operationally tested on an annual basis. The accuracy of simulator computed values must be determined over three points across the acceptable parameter range (i.e. low, middle, high). Simulator design control is initially based on predicted plant performance (if the nuclear reactor is new) until eighteen months of hard plant performance data is available. Design control is then based on the last eighteen months of historical data. Student feedback is extensively used when simulator modifications are considered. Modifications to the simulator must be made within twelve months of initialization.

To establish simulator credibility, simulator output response is compared to the actual power plant response under the same conditions. To evaluate simulator response for unexperienced actual power plant conditions (i.e. runaway reactor approaching meltdown), the simulator's generated output is compared to the best engineering estimates provided by nuclear experts (American National Standard, 1985).

b. Safety Parameter Display Systems

The safety panel display system (SPDS) is a priority safety improvement on nuclear reactors. The SPDS displays critical parameters that indicate proper functioning of nuclear power plant systems and initiates a warning to power plant personnel that parameters are out of bounds (Straker, 1981). The verification/validation plan for the SPDS was based upon balancing the depth of the verification/validation effort against the quality control requirements for the system. Because the

SPDS is not as critical as a nuclear reactor protection/containment system, it would not be subjected to the same level of verification/validation effort. A verification/validation plan for SPDS was developed for the Nuclear Safety Analysis Center by Science Systems, Inc. in 1981. This plan underscored the importance of incorporating the reaction of licensed power plant operators to the SPDS under test. It highlighted the importance of incorporating verification/validation into the entire development cycle of the SPDS rather than executing it as an afterthought to a finished product. To prevent bias, it is necessary to separate the verification/validation of the SPDS from the developer who has a vested interest in verification/validation performance results. Documentation of the entire verification/validation effort was deemed to be critical for traceability of all work done (avoid duplicated effort), identification of strengths and weaknesses, and continuity between users. Verification/validation was considered to be a five stage process:

(1) System requirements review: to determine if requirements are correct, complete, consistent, feasible, and testable.

(2) Design review: check hardware and software to ensure that requirements for the system are correctly met.

(3) Validation test: after developer certification, conduct an independent verification/validation by the user to verify developer claims.

(4) Field verification test: ensure that the validated system is correctly installed in the field where it will ultimately be used.

(5) Validation report: the documentation of all verification/-validation activities to serve as a reference for future review, modification, and study (new users) of the system (Straker, 1981).

c. Software reliability

As nuclear reactor computer control systems are modified or upgraded, it remains critically important for the computer software to be absolutely reliable, particularly those computer codes and routines that manage vital aspects of plant operation such as emergency cooling procedures. A survey of software assurance methods for the U. S. Nuclear Regulatory Commission, prepared by EG&G Idaho, Inc. in 1981 identified the need to conduct verification/validation of computer coding commensurate with safety requirements and the quality of software needed (Smith, 1981). Redundant coding can be used to increase reliability in critical applications. With respect to software testing, the initial look should focus on how well the code is written. Structured programming techniques should have been used with documentation built into the code to describe key variables, subroutines and functions. The code should be readable, maintainable, and consistent. Automated code checking programs are especially helpful in tracing model execution through the code. Deskchecks and peer reviews by programmers are traditional code evaluation techniques. It is critical that the testing of the code be planned thoroughly and conducted by experts who specialize in testing. Testing should be conducted to specific standards and all results documented.

d. Verification/Validation at Plant E. I. Hatch

The Plant Hatch training simulator was developed in accordance with the standards specified in the American National Standard

ANSI/ANS-3.5-1985. The training simulator models the nuclear power reactor at the site. During the simulator's acceptance testing, nuclear engineers and simulator experts compared the output of the simulator to the actual readings obtained from the plant. The physical mechanics of the plant were modeled against plant blue prints and their specifications for valve tolerances, stroke times, etc. The simulator's power output was checked against the actual power output data generated by the plant. Hard data was available from the design specifications of the plant and historical power output data to ensure that the simulator accurately portrayed the physical nature of the plant. Due to the lack of empirical data for actual plant performance in extreme conditions (i.e. runaway core leading to potential meltdown), the simulation output for accidents, emergencies, and some extreme transient conditions were modeled after expert opinions and conservative engineering estimates. Internal plant documents govern the modification and testing of the simulator (Simulator Configuration Control, 1987). A Simulator Modification Review Committee (SMRC), staffed by plant personnel, continuously scrutinizes the simulator for accurate plant simulation. An in-house Configuration Control Program ensures that design changes and modifications to the plant are added to the simulator in a well orchestrated and responsible fashion. These changes may include changes to simulator initial conditions, changes to the reference plant data base, changes to correct operator observed simulation discrepancies, enhanced simulator capabilities, and hardware replacements. Plant personnel who identify simulator output that does not mirror actual plant performance submit specific discrepancy reports that describe the observed discrepancy in detail, including the actions or events that led to its occurrence.

Simulation runs are repeated under the same conditions in an attempt to isolate the cause of the discrepancy. The simulator data base, software, or hardware is then modified accordingly (Simulator Configuration Control, 1987).

2. Inferences

a. Establish documentation that specifies standards and procedures for the verification/validation of military models in use or to be developed at CAA. The documentation should establish a clear and precise framework that indicates the authority and responsibility of key personnel in the verification/validation process. It should outline the specific steps, tasks and objectives to be accomplished in the verification/validation process.

b. When model output cannot be compared to historical data, consider the model to be credible if a highly experienced team composed of system experts cannot distinguish between simulator output and actual system performance (a Turing test).

c. Develop and install the additional code to incorporate user intervention into the simulation model. Determine if freeze simulation, backtrack, and snapshot capabilities would enhance the simulation model's usefulness. Modify the model to generate periodic output of key parameters, variables, and conditions in hard copy form. Incorporate key parameter annunciators into the model that signal potential out of bounds errors in model parameters or variables during simulation. This allows

the user to identify puzzling values and focus on particular modules or phases of the simulation where validity may be in doubt.

d. When validating the model, conduct simulation runs with key model parameters and variables initiated over a wide range of their permissible values to ensure that corresponding output remains within credible limits (extreme condition tests).

e. Continuously monitor and review user critiques of the model for their opinions of model accuracy. Users at the grassroots level may identify particular simulation runs that appear to be unreliable under certain initializing conditions and data sets. These initialization conditions and data sets may have escaped scrutiny during earlier initial verification/validation efforts.

f. Ensure that model modifications are thoroughly tested before their implementation is made. Identify a strategy and develop internal procedures for the conduct of this testing.

g. When a historical data base of system performance is not available for comparison to model output (i.e. general war in Central Europe in the 1990's), generate a "best estimate" data base of system response from similar historical events, the physical limitations and capacities of equipment, the results of similar simulations, military exercises (i.e., REFORGER), and the experience of military experts. Run a full scale simulation of the model with initial values of parameters and variables at their most likely values and compare output results to

the "best estimate" response. Consider the model to be credible if the model output closely resembles the "best estimate" response.

h. Determine the level of verification/validation necessary and acceptable for the particular model under review. Critical portions of the model may require extensive and exhaustive verification/validation effort. Other modules or subroutines may require less effort. Make this analysis prior to initiating actual verification/validation of the model so that resources, time and effort may be most economically expended.

i. Consider structuring the verification/validation of the model around the following process:

(1) Model requirements review: ensure that the model, as described by specified requirements, models the appropriate problem. Thoroughly review and specify each of these performance requirements and operational capabilities in a reference document. Ensure that these requirements/capabilities can be adequately tested by the verification/validation team assembled. Make each performance requirement or capability a verification/validation mission. Develop a "roadmap" document that matches every verification/validation mission to the requirements document by allocating personnel, equipment, suspense times and test standards for specific verification/validation missions to their counterparts in the requirements document. Establish a tracking system for all verification/validation missions by developing a matrix which establishes suspenses and matches all verification/validation missions against the model requirements. Use the matrix to track

requirements through the design review phase into the validation testing phase of the verification/validation process. The specific requirements and capabilities of the model (verification/validation missions) make up the rows of the matrix while the columns identify the portion of the requirements document that addresses the requirement or capability.

(2) Design review: Review the current necessity for and needs of the particular simulation model. Examine the hardware and software specifications of the model to determine if they need modification to support current requirements. Ensure that CAA resources can adequately meet the updated requirements in the requirements document.

(3) Validation test: This portion of the verification/validation process consists of two parts, the test plan development and test execution and results analysis. The matrix developed during the model requirements review is used to track the specific verification/validation missions to be tested. Evaluation techniques, strategies and standards are developed for each mission. Personnel and time are allocated to each mission. The tests are then conducted in accordance with the test plan. Finally, the results are analyzed.

(4) Validation report: Document all verification/validation activities and maintain for reference. This documentation will assist analysts who may modify or troubleshoot the model in the future.

j. Examine all of the code prior to initiating verification/-validation and categorize modules, subroutines, and functions as to their

importance in determining model reliability. Allocate verification/-validation efforts and resources accordingly.

k. Utilize redundant coding in critical portions of the model.

l. Examine the code to ensure that it is understandable, accurate, and complete. Consider subroutines, functions and modules to be "black boxes". Once they are certified as reliable they can be exempt from further verification/validation (Smith, 1981).

m. Utilize automated code checking software to run traces through the code. Select automatic code checking software that will provide written documentation of trace results for later reference.

n. Develop a verification/validation plan for software prior to starting the work. Ensure that each requirement is conducted to a specific standard. Utilize experts to conduct the code checks.

o. Review the military model by experts. Ensure that the simulation data base is modeled after available hard empirical data as much as possible (i.e. initial conditions for the quantity of diesel fuel in the basic load of a tank battalion should be checked against current MTOE. Subroutines governing fuel consumption should be patterned after historical usage reports experienced by tank battalions). Considerable effort should be expended to ensure that the model's initial conditions are accurately maintained in the data base.

p. Compare simulation results to available real world data. Results may be compared to similar simulation model results, military exercises (i.e. REFORGER) and the "best estimates" of recognized experts.

q. Develop a mechanism for documenting discrepancies in the simulation model. Be alert to the experiences of the user. Provide a means for the user to flag discrepancies in a timely fashion. User's must be trained to document the initializing conditions of model runs. Discrepancies in model behavior or output can then be examined through diagnostic runs under the same initializing conditions. This will assist analysts in corrective troubleshooting.

References:

American National Standard Nuclear Power Plant Simulators for Use in Operator Training (1985), American Nuclear Society, La Grange Park, Illinois, pp. 1-6.

Simulator Configuration Control (1987), Departmental Instruction Procedure, Document Number DI-TRN-37-0787N, Georgia Power Company, pp. 2-10.

Smith, R. (1981), Survey of Software Assurance Methods, EG&G Idaho, Inc., Idaho Falls, Idaho, 1981, pp. 1-40.

Straker, E. A. (1981), Verification and Validation for Safety Parameter Display Systems, Science Applications, Inc., La Jolla, California.

D. Large system: TAC Thunder

1. Discussion

TAC Thunder is a theater level combat simulation model simulating the major aspects of air and ground combat, deployment and resupply, airlift and sealift. The model can simulate a full scale conventional war anywhere in the world merely by the player changing the initial conditions in the model data base. Separate data bases exist for red and blue forces and simulate their own doctrine, strategy, and tactics. It is a large model, consisting of 118,000 lines of SIMSCRIPT II.5 code. TAC Thunder was converted into an interactive, joint service theater war game to satisfy a pressing need for training American commanders and their staffs at the operational level of war (der Boghossian, 1988). The wargame facilitates training in command and control in a joint environment without the need for large numbers of players at subordinate levels of command. The conversion process of the TAC Thunder simulation model to a wargame reveals effective verification/validation techniques that can be applied by the CAA modeler.

The original TAC Thunder simulation model was converted to an interactive wargame in seven months. The requirements for interaction were compared against the capabilities of the existing model and modules requiring modification were identified. This process was simplified by the fact that TAC Thunder possesses a highly modularized structure (TAC Thunder has over 600 modules with naming conventions followed to categorize subroutines and modules by functional groupings, is well documented, and has very readable code). Modules that previously updated parameters and made decisions "through the computer" were modified to

allow the players to modify these parameters and decisions. TAC Thunder was modified to allow the user to intervene at various levels within the simulation. The user can intervene to change such things as operational boundaries, assign aircraft to targets, assign target priorities, and allocate forces to commanders. Controllers can intervene to modify such things as POL available at an airbase, characteristics of aircraft, characteristics of munitions, weather, etc. TAC Thunder provides back up and look ahead capability during simulation and produces hard copy data files and reports (der Boghossian, 1988).

The modification of TAC Thunder started with a careful review of the requirements specified for the interactive game. Tough questions were asked. What do we want the interactive model to do? The existing model was examined from the top down to ensure that it possessed the capability to meet all requirements. The model was then systematically examined from the bottom up and broken into tractable pieces for modification. Modules requiring modification were thoroughly walked through and modules requiring change identified. After the code changes were made, driver programs and data sets were developed to test them for validity. Test runs and walkthroughs established the validity of changes. Data bases were then constructed to test additional sections of the model for validity. The result of this bottom up process was the eventual run of the entire simulation against prepared data sets for overall model validation. Simulation performance was compared to the specifications in the requirements document. Differences were identified, corrected, and checked again. Personnel who wrote the code were not included in the testing phase. This eliminated a source of potential bias. Throughout the modification process, verification was built in and conducted as the

work proceeded. Modules were verified on an individual basis first, then as part of a larger whole. The old adage, "if you don't have the time to do it right the first time, how will you do it right the second time?", was particularly applicable here.

2. Inferences

a. Before CAA invests the resources to verify and validate an existing model, it must ask tough questions. Why did we build this model? Will another model do a better job? What are the current requirements for the model? Are the current requirements much different from those when the model was originally built? Have assumptions changed? Have new technologies in weapons and equipment made the model useless? Have changes in strategy and tactics made the existing model inapplicable or grossly inaccurate? Can this model be modified to meet the new requirements or should we start from scratch? What must this model do today? Concrete answers must be given for each question. Requirements for the model today must be specified. The existing model must be examined carefully to ensure it meets all of today's requirements. If it does not, it must be modified. The modification process must be systematic and thought through clearly. The model must be reviewed from the top down to ensure it has the capacity to fulfill all of today's requirements.

b. The model is then broken into manageable, tractable parts. The model is checked from the bottom up to ensure modules and subroutines are constructed to support today's requirements. Data bases are thoroughly

reviewed for accuracy to ensure they reflect current equipment characteristics, force structure, equipment capabilities, tactical deployment, tactical employment, etc. Computer code is walked through and test runs on modules and subroutines verify changes as they are made. Verified portions of the model are then combined into larger sections for validation runs and evaluation. All validation testing is done with data bases updated to reflect current values as specified above.

Reference:

der Boghossian, Z. C. (1988), TAC THUNDER - Building a Wargame, CACI Products Company, Arlington, VA.

E. Large system: Strategic Defense System

1. Discussion

The "Stars Wars" research program is an example of a large system for which hard historical data is generally not available for comparison with simulation results to determine validity. Thus, it is an excellent analogy to many military models. Verification/validation techniques utilized in SDS research is directly applicable to the military models in which CAA is interested. Many components of the SDS system are still in the early conceptual design stages of development. Physical and operational characteristics of the systems have not been finalized. Physical models of the systems do not exist. Consequently, credible simulation models must be developed to provide answers concerning SDS design questions and system performance estimates as a part of system development.

The MITRE corporation has prepared a draft document entitled "Guidelines for Evaluation of SDS Simulation Models Used at the National Test Bed" (Gados, 1988) that addresses these verification/validation techniques. The document stresses the need to organize all simulation resources and activities for maximum efficiency. Top down simulation requirements analysis is heralded as is verification/validation conducted by an independent evaluation team. The need to document all simulation evaluation is stressed. The concept of verification and validation of SDI systems not yet in physical existence is analogous to the war not yet fought:

> A model of a complex, real-world SDS is at best only an approximation to the actual system. Complex simulation models cannot be designated as being absolutely valid over the complete domain of all possible uses or applications. This is because it is too costly or time consuming to test and evaluate all possible conditions of the model's uses. Since the SDS system currently does not exist, it is not possible to test the actual system's behavior and compare it with the models' behaviors to determine their accuracy. SDS simulation models cannot be considered completely valid or invalid. Instead, the establishment of simulation credibility should be considered an evolutionary process of obtaining sufficient confidence in the models' behaviors and their capabilities to provide quantitative answers to structured posed problems (Gados, 1988).

In an analogous manner, CAA should quantify the overall validity of military models in terms of their abilities to provide believable answers to specific questions.

The overall process of model validation/verification described by the MITRE corporation consists of reviewing simulation development, conceptual model validity, computer model (software) verification, operational validity, data validity, and internal security verification.

Reviewing simulation development consists of assessing the current state of the simulation model. Conceptual model validity is determined through statistical analysis and mathematical analysis to see if the model is a reasonable representation of SDS concepts. Computer model verification determines if the computer programming that implements the conceptual model is correct. It is performed through computer-aided design, walkthroughs, and preliminary testing. Operational validity of a simulation (currently unobservable in the physical world) can be inferred by comparing the simulation model to other credible models already in use. Data validity consists of examining input data by expert reviewers to ensure data correctness. This data can be verified by comparison to

the data bases of other credible models. Internal security verification suggests placing a "lock" on verified code to ensure it is not tampered or modified by mistake at a later date (Gados, 1988).

It may be impossible to prove that a simulation model produces "correct,"perfect," or "absolutely reliable" output. One approach may be to accept the output as being valid if it cannot be proven to be incorrect after analysis by experts.

An important check is to ensure that the requirements of the model are completely specified. If the model lacks certain capabilities important to the modeled system or to the user then it should be considered not valid until corrections are made.

Gados outlines a number of useful approaches that can be applied to existing military models for validation and verification:

Review the Simulation's Development:

(1) Initiate the review. The purpose of the review is to identify the overall status of the simulation model. The simulation model is examined and a determination of its ability to simulate the system assessed. It is important to identify any important characteristics of the system that are not adequately addressed in the current state of the simulation model.

(2) Problem definition. Experienced personnel familiar with the intended use of the simulation examine the "problem" for which the simulation model was designed to ensure that the model is adequately specified. Only through thorough specification of the problem's (system's) requirements can the simulation model be later judged for its ability to correctly model the problem.

(3) Model specification. The conceptual model of the problem is thoroughly reviewed to ensure that it is detailed, specific, and accurate. The model specification must be complete if the simulation model is to be credible. The credibility of the simulation model is judged on its ability to provide concrete answers to specific system questions.

(4) Software engineering. The software is examined to determine how well it was constructed and implemented. Good software design calls for methodical implementation and specific adherence to standards.

b. Techniques for conceptual model validation:

The conceptual model is tested to determine if algorithms, assumptions, concepts and ground rules are accurate. The level of detail and fidelity in the model is assessed. Model inputs, outputs, range of simulation, methods of engagement, threat characteristics, and operating environment of the model are assessed. Techniques used in the assessment include subjective analysis, historical review, empirical testing, and logic traces.

(1) Design credibility. The very design of the model is examined to determine if it is robust, testable, affordable, effective, maintainable, and useful. The model should be capable of accepting expansion and growth if necessary.

(a) Model assumptions and ground rules. Assumptions and ground rules used in the development of model algorithms are examined for accuracy. Internal data values in the simulation model are checked for accuracy. Weapons details are examined for accuracy (weapon type, ranges, ammunition types, probabilities of kill, reaction

times, target engagement times, target acquisition times, survivability, movement rates, etc.). Battle management, command, control, and communications are examined for accurate portrayal of friendly and threat doctrine and capability.

(b) Component level of detail. All modules, subroutines, and components of the simulation model are examined to ensure that they provide the correct amount of resolution and detail. Algorithms or subroutines that interface with each other are examined to ensure that they have compatible levels of detail.

(2) Concept validity. The concepts used to develop algorithms, modules, and subroutines are examined for accuracy.

(a) Historical derivation. Review the historical development of the model to ensure it was based upon sound principles and decisions. If algorithms, modules, subroutines, or sections of code were obtained from previous models (forerunners of the current model) their credibility should be judged.

(b) Logic traces. Traces are made through the model's algorithms and modules to assess adherence to the conceptual model and to the operational nature of the modeled system. Specific model entities (i.e., a target acquisition radar) are traced through the model to ensure their behavior in the model adheres reasonably well to actual characteristics.

(c) Face validity. Experts familiar with the system examine the model and subjectively determine its accuracy and worth. Questions they raise should be addressed through further model evaluation.

c. Techniques for software validation:

(1) Software development standards. Examine the underlying standards used to construct the software. If a particular military standard was used (such as Mil. Std. 2167), it is a straight forward process to walk through the code and examine it for compliance.

(2) Internal software testing. Test portions or modules of code with driver programs to ensure they function as intended. The driver program should force the code to execute through all possible decision steps and routes.

(3) Correctness proofs. A mathematical proof may be constructed by an expert to verify the correctness of a module or portion of code.

(4) Automatic code checkers. Use automated code checking programs to identify uninitialized variables, inconsistent declarations, incomplete statements, infinite loops, unused sections of code, etc.

d. Techniques for operational validity:

(1) Input-output relationships. Search for and examine the causal relationships between model output and internal values or inputs. Systematically determine and categorize the relationship between input parameters and output results. Determine if unrelated and independent inputs cause changes in outputs that should not occur in the "real world."

(2) Event validity. Review a simulation run by experts to determine if the sequence of events generated by the model are believable and occur in a logical sequence or pattern.

(3) Turing tests. Execute a simulation with carefully controlled input data, then have the output reviewed by experts. If the expert can

readily categorize the simulation output from real world occurrence or estimated real results then the simulation may not be credible.

(4) Delphi technique. A panel of recognized experts reviews the output of the simulation model. The panel's consensus of opinion establishes the credibility of the model.

(5) Demonstration techniques. Execute the model to perform with a selected data base under a certain scenario for which it was designed. Examine how well the model performs the simulation.

(a) Simplified assumption testing. Run the model under simplified assumptions which are known to be true and examine the output for correct response. Remove modules or portions of code and examine response. Vary one input parameter or set of parameters that effect a specific assumption of the model, then examine the model output to see if the correct response was generated.

(b) Animation. Utilize graphics to examine the model's objects and events to see how they move through time. Use currently available automated tools to construct the graphical displays of simulation results.

(c) Predictive validation. If possible, construct physical models of the entire system or portions of the system. Compare the performance of the physical model to simulation results under the same initialization conditions.

(6) Analytic techniques. Use quantitative statistical tools (hypothesis testing, analysis of variance, analysis of covariance, multivariate techniques, regression analysis, response surface techniques, nonparametric techniques, discriminant analysis) to compare simulation results with external standards.

(a) Compare overall results to a standard. Develop standards for expected simulation response, under various initialization conditions, by collecting data from actual tests, historical records, or best estimates of the actual system. Then compare the simulation output, under each set of initialization conditions, to the standard. Define each comparison to be made between simulation output and real system performance. Generate a set of initial conditions for each comparison. Identify statistical requirements (confidence levels, significance levels, etc.) and determine the number of replications to be made. Then utilize the statistical techniques mentioned above to analyze the relationship between model output and the "real" standard. This analysis requires a dedicated research team and is resource intensive, time consuming, and often difficult.

(b) Comparison to test data. If specific test data is available for the real system the model can be initiated to reproduce, as closely as possible, the conditions under which the test data was obtained. Model output can then be compared to specific test data by statistical methods.

(c) Sensitivity analysis. Modify input values or parameters over their entire range and examine model output. Identify those inputs that cause significant changes to output response and examine them carefully for accuracy.

(d) Predictive validation. Produce a prediction of simulation results by scientific analysis, engineering analysis, other simulation, or actual system test. Then compare simulation output to the predicted response.

e. Techniques for data validity:

(1) History. Review the history of input data (where it was obtained and who obtained it) to access its validity for the model.

(2) Internal consistency. Check all input data against a range of acceptable values. During simulation, monitor them to ensure they do not go out of bounds.

(3) Accuracy of implementation. Examine input data constants for accuracy. Ensure that input random variables are of the correct probability distribution (Poisson, exponential, etc.).

(a) Portrayal of constants. Ensure constant values are installed correctly, addressed properly, and remain unchanged through the simulation.

(b) Justification of distributional form. Utilize statistical goodness-of-fit tests to verify the input distributions of random variables.

2. Inferences

a. Accept the fact that it will be impossible to completely validate and verify a model in use at CAA with 100% accuracy. Instead, determine a level of accuracy that satisfies the needs of CAA and produces results accurate enough to support decision making at Army levels.

b. Review the model outputs to ensure that they will generate all data needed by CAA. Then utilize the six step process of reviewing

simulation development, conceptual model validity, computer model verification, operational validity, data validity, and internal security verification to systematically review the model.

c. Determine the needs of models in use at CAA. Evaluate the techniques of reviewing simulation development, conceptual model validity, software validation, operational validity, and data validity described above for applicability to CAA models. Once the applicability of techniques is identified, determine the time and resources available at CAA to apply them. Select those techniques that support CAA's acceptable level of verification/validation while satisfying time, budget, and other external constraints.

Reference:

Gados, R. G. (1988), Guidelines for Evaluation of SDS Simulation Models Used at the National Testbed: A Preliminary Report of the Simulation Evaluation Methodology Subgroup, The MITRE Corporation, Falcon Air Force Station, CO.

F. Large system: National Airspace System

1. Discussion

The MITRE corporation is currently validating a National Airspace Simulation model for the Federal Aviation Administration (FAA). This model describes, in an aggregate manner, the effects of many interactions of the airspace system without so much detail that data collection and model run times would be prohibitive (Cheslow, 1988). Important aspects of the system have been incorporated into the model without including detailed representation of the activities of individual airports or air traffic controllers. The model is being developed in two phases. The first phase, upon which the following discussion is based, is a deterministic simulation of air traffic using 58 airports and their local airspace. The second phase will include stochastic elements and expanded airspace coverage. Major inputs to this model include scheduled departures, scheduled arrivals, unscheduled business and military flights, service rates at airports for arrivals and departures, service times, weather patterns and its effect on service rates and flights. Validation of this model rests in determining accuracy in an aggregate sense. The validation process consists of four parts: conceptual model validation, input data validation, computerized model correctness checks, and operational validation (Cheslow, 1988).

The phase one testing of the airspace system has been completed. Results of the testing were presented at the 1988 Winter Simulation Conference in San Diego California.

Conceptual model validation tested the logical structure and reasonableness of the model to ensure that the model was consistent with

physical and engineering constraints such as airport locations, aircraft speeds, distances between airports, and air traffic control procedures. Due to the aggregate nature of the model, it was important for this model to experience little change in output when simplifications to the model were made.

Input data validation focused on ensuring that data used to develop parameters and input variables were accurate, reasonable, and consistent. Data was obtained from three different reporting agencies. Airport records were used extensively. Data was checked against the physical abilities of the system (i.e., aircraft handling rates at a particular airport) to ensure inaccuracies were not built into the model.

Computerized model verification focused on checking the computer code and system design for errors. Outside experts reviewed the code and conducted traces on computer code to ensure accuracy.

Operational validation focused on ensuring that the model produced valid output. This process was simplified by the fact that accurate historical data exists for comparison. Outputs were checked for boundedness to ensure that negative values were not generated (i.e., for such outputs as service times). Degeneracy checks were made to ensure that changes in the input data produced appropriate changes in the output (i.e., delays at airports did not increase when actual aircraft loads were reduced).

Extreme input behavior was examined to ensure that the model responded as expected in the limit. For example, aircraft arrival delays were eventually eliminated by the introduction of increased airport capacity and availability. Graphical presentation of model outputs were

examined by experts for anomalies (a picture is worth a thousand words and provided ready evidence of model behavior).

Statistical checks (calibration and sensitivity analyses) were conducted as a part of operational validation. Calibration consisted of modifying the model to ensure that model outputs matched prescribed values. Mismatches between outputs and prescribed values were closely examined to determine patterns. The causes for these patterns included errors in input data, algorithms, computer code, and the calibration data set. Sensitivity analysis examined the differences in model output generated by changes in input parameters before and after changes were made. Precise statistical tests were not employed because of the possibility of autocorrelation in real world experience and model output. Statistical tests were discarded in favor of an accuracy criterion. This criterion established the requirement for the model to be as accurate as the database it was compared against and accurate enough to produce output useful in the analysis of the issues for which the model was designed. Twenty to forty percent uncertainty was not considered to be unreasonable for this particular model. Since a model can never be considered absolutely valid for all conditions, the goal of validation was to produce a model "good enough" for its intended use (Cheslow, 1988).

Validation measures were constructed to allow easy comparison of model output to existing historical data. For example, if historical data allows the accurate computation of average delay per aircraft per airport, then model output generated this data. Output measures unsupported by a historical data base were not utilized. Unused output

measures not useful to the user were not generated, even if their generation was simple and straightforward.

Significant model improvements resulted in the adjustments made during early quantitative calibration. The matching of outputs to prescribed values led to many refinements in algorithms, input parameters, and computational formulae.

2. Inferences

a. Structure the validation process at CAA around the four steps: conceptual model validation, input data validation, computerized model verification, and operational validation.

b. Consider and examine each specific input that goes into a military model at CAA. Ensure that the input data base accurately reflects the physical and/or engineering limitations of military equipment, personnel, and tactics. For example, if the model includes the shipping of war materiel in existing merchant fleet vessels, ensure that cargo capacities, steaming times, load times, and unload times for each class of vessel are accurately entered into the input data base.

c. Gather historical data to construct an "output data base" for comparison to simulation model output. If a "real world" data base for system performance does not exist, use expert opinion, engineering best estimates, inference from similar historical events or systems, recent modern conflicts (Falklands, Israeli intervention in Lebanon, etc.), and military exercises (REFORGER, RED FLAG, etc.) to construct the output

data base. Compare simulation results to this output data base. Based upon differences, develop prescribed values for what the model output should look like and use them to "calibrate" the model. Incorporate sensitivity analysis into this process. Modify variables, mathematical formulae, variable relationships, modules, and subroutines to "tweak" model output.

d. Establish reasonable bounds for model outputs and internal parameters. Check the model to ensure the bounds are not exceeded. If possible, build a self-check algorithm into the model to signal out-of-bounds errors.

e. Conduct degeneracy checks on key parameters and algorithms to ensure that changes in input variables lead to changes in output in the right direction (i.e., increasing the number of tanks in Blue tank battalions should result in increased kills against Red forces. If model output reflects a decrease, then troubleshooting is needed in particular subroutines or algorithms).

f. Develop an accuracy criterion for each CAA model. Careful consideration is given to the required level of accuracy needed in model output to support decision making. The accuracy criterion for each model is based upon the results of this analysis. The model can only be as accurate as the "real world" data base (historical or contrived) against which it is compared. The simulation model only needs to be accurate enough to instill confidence in the decision maker or force planner who uses it.

g. Examine model output to ensure that it generates the information necessary for informed decision making or analysis by CAA. The best output is that which provides the data necessary for informed decision making and which can be compared against an accurate historical database or quantified expert opinion, best guess, or engineering estimate.

Reference:

Cheslow, M (1988), Validation of a Simulation Model of the National Airspace System, 1988 Winter Simulation Conference Proceedings, San Diego, pp. 791-794.

G. Large System: Marshall Space Flight Center

1. Discussion

The Marshall Space Flight Center (MSFC) in Huntsville, Alabama performs research and development for launching and operating spacecraft for NASA. The Principal Investigator initiated a series of telephone calls with various researchers at the MSFC in the Summer of 1988 to determine how verification/validation activities are conducted.

a. Software requirements.

Discussion with Ken Williamson (Systems Software Branch) revealed that, in general, software development and software verification occur simultaneously and evolves from an initial requirements specification to a more detailed requirements specification as time progresses. All input and output requirements are completely specified. Software is tested against specific standards. Extreme condition tests are used wherever possible to validate the software. Problems with software verification are documented as they occur. Strict configuration control of the software development process is considered to be very important. "Pride in one's module" is stressed among all programmers and analysts.

b. General verification and validation procedures.

Conversations with Pat Vallely (Structures and Dynamics Lab) revealed that definitive guidelines for verification/validation at MSFC are not specified. In general, the temptation to create too large a model is avoided. (Orbiter engine models, for example, require 40,000 to 50,000 lines of FORTRAN. Models this large should not be made more

cumbersome by adding capabilities not specified in current requirements). During simulation development, non-essential parameters are eliminated and the model is reduced as much as possible, while ensuring that model outputs satisfy all performance requirements and specifications. The inputs and outputs for each module are thoroughly reviewed. Connections (data transfer) and feedback between modules are traced for logical flow and accuracy. One useful tool for verification is a matrix representation to portray COMMON block statements in the model. The matrix allows the easy cross-reference of specific COMMON block variables to the modules in which they are referenced. Errors in variable declaration and type can be easily identified with this procedure. Discussion with Dave Aichele (Software and Data Management Division) identified commonly used procedures for verification/validation in the Information and Electronics Lab. Simulation requirements are thoroughly reviewed and checked against the simulation model. Test runs are made with median input values for all parameters. Simulation output is compared to the predicted response. Test runs are then made with extreme values for parameters to evaluate corresponding simulation response. Truth data (expected values) from earlier, credible models are often compared to simulation output.

c. Space Telescope Simulator.

The Space Telescope Simulator was developed to train astronauts on the ground to operate the space telescope from an orbiting space shuttle. Prior to developing the simulation, the use of an existing simulation model was considered. This alternative was discarded due to the potential inability of certain existing modules to correctly represent the present system. Additionally, the earlier simulation was written in a different computer language and translation to the desired language was

considered to be too difficult. The original concept for simulator testing required the MSFC to develop the model and portray the actual system during testing. The Goddard Space Flight Center (GSFC) would serve as ground control, sending electronic commands to MSFC to perform various space telescope functions and exercise its capabilities. MSFC would then simulate the transmission of telemetry data back to GSFC. In actuality, the simulator was developed and ready for testing at MSFC prior to the completion of operations procedures at GSFC. Consequently, MSFC developed their own testing procedures for the simulator and constructed operational scripts to play through the system. Conversations with Ellen Williams revealed that reacting to changing requirements was the biggest problem encountered in the simulator verification/validation process. During construction of the simulation model it was difficult to initially obtain all requirements. As time progressed, more and more fidelity in the model was required (to provide more detailed training for astronauts). Once the actual system was constructed, the simulation model had to be modified to adapt to design changes in the actual system. The simulation model was validated, in part, by comparing simulation output to actual performance of the telescope system in space.

d. Payload Crew Training Complex (SPACELAB) Simulator.

This simulation model was developed to simulate the operation of the SPACELAB experimental station (a self-contained scientific laboratory for use in the space shuttle). The simulation model played the part of the actual system in space and ground control. It simulated the transmission of commands from ground control, the performance of the

experimental stations, the keystrokes entered by astronauts in the system computers to direct the experiments, and the transmission of data back to Earth. The model was constructed with a large number of modules. Discussion with Steve Puritan (Systems Software Branch) revealed that each module was developed and tested individually prior to testing in an aggregate sense. This verification involved running the modules through a number of orbiter profiles (scenarios). Verified modules were then evaluated together as part of a larger system, gradually moving to an evaluation of the entire simulation model. Difficulties were encountered in taking a verified module and verifying/validating its performance as part of the larger system. These problems were caused primarily by the difficulty in tracing the changes and modifications of various parameters and variables as they passed between modules. Performance requirements for the simulation model did not demand high fidelity. Output had to be reasonable. Unrealistic data generated by the model forced additional verification/validation to identify the cause. The final step in the verification/validation process was an acceptance test. Model output was examined for a given set of prescribed scenarios. Experts involved in the project examined the output data. Their consensus of opinion on the accuracy of the output data established the credibility of the simulation (Delphi technique).

2. Inferences

a. Insure that all performance requirements and desired capabilities for a simulation model are completely specified and understood by CAA analysts prior to the start of verification/-validation. Avoid changing verification/validation requirements once the

process has started. Specify the required fidelity of the model early in the verification/validation process. Avoid the temptation to add capabilities to a model that exceed performance or documentation requirements.

b. Test software against specified standards. These standards should be specified prior to the start of verification/validation.

c. Utilize extreme condition tests whenever possible to evaluate the performance of algorithms and modules. Simulation runs should be conducted with all parameters and variables initialized across a wide spectrum of permissible values (low, middle, and high). Utilize median input values for parameters and compare simulation response to predicted response. Utilize truth data (output response known to be true for a given set of initial conditions) as a comparison data base for simulation output.

d. Document all verification/validation results throughout the model evaluation process.

e. Verify the separate modules of a model first. Then aggregate the modules and test them as a larger whole. Consider the incorporation of existing, verified modules into a simulation model. These modules should be used only if their performance and response in the new model is thoroughly known.

f. Utilize matrix representations of parameters, variables, and their declarations to trace interface between modules.

g. Develop reasonable performance requirements for the model. Consider the model output to be valid if unreasonable data values are not produced and output is considered to be reasonable by a consensus of expert opinion (Delphi technique).

H. Conclusions

1. Establish a standard document at CAA that specifies standards and procedures for the verification/validation of military models in use or to be developed at CAA. This document should establish a clear and precise framework that delineates the authority and responsibility of key personnel in the verification/validation process. It should outline the specific steps, tasks and objectives to be accomplished in the verification/validation process. Develop internal procedures for documenting discrepancies observed in CAA simulation models. Be alert for discrepancies observed by the users. Provide a means for the users to flag discrepancies in a timely fashion. User's must be trained to document the initial conditions of the model that led to the discrepancy in order that future diagnostic runs may duplicate the problem for corrective troubleshooting.

2. Ensure that model modifications are thoroughly tested before their implementation is made. Document all modifications. Identify a strategy and develop internal procedures for the conduct of this testing and documentation.

3. Accept the fact that it will be impossible to completely validate and verify a model in use at CAA with 100% accuracy. Instead, determine a level of accuracy that satisfies the needs of CAA and produces results accurate enough to support decision making at Army levels. Develop an "accuracy criterion" for the results generated by

each CAA model. This criterion is established after careful consideration of the accuracy required in output to support the decision making process the model was designed to support. The model can only be as accurate as the "real world" data base (historical or contrived) against which it is compared. The simulation model only needs to be accurate enough to instill confidence in the decision maker or force planner who uses it.

4. Determine the level of verification/validation necessary and acceptable for the particular model under review. Critical portions of the model may require extensive and exhaustive verification/validation effort. Other modules or subroutines may require less effort. Conduct this analysis prior to initiating actual verification/validation of the model so that resources, time and effort may be most economically planned and expended.

5. Before CAA invests the resources to verify and validate an existing model, it must address tough questions. Why did we build this model? Will another model do a better job? What are the current requirements for the model? Are the current requirements different from those when the model was originally built? Have assumptions changed? Have new technologies in weapons and equipment made the model useless? Have changes in strategy and tactics made the existing model inapplicable or grossly inaccurate? Can this model be modified to meet the new requirements or should we start from scratch? What must this model do today? Concrete answers must be given for each question. Requirements for the model today must be specified. The existing model must be

examined carefully to ensure it meets all of today's requirements. If it does not, it must be modified. The modification process must be systematic and thought through clearly. The model must be reviewed from the top down to ensure that it has the capacity to fulfill all of today's requirements.

6. Determine the necessity for and needs of the particular simulation model. Examine the hardware and software specifications of the model to determine if they need modification to support current requirements. Perhaps new software is needed to allow interaction with other models. Perhaps faster computers are needed. Ensure that CAA resources can adequately meet the updated requirements of the model.

7. Review the model outputs to ensure that they will generate all relevant information needed by CAA to support current analysis. Examine model output to ensure that it generates the information necessary for informed decision making or analysis by CAA. The best output is that which provides the data necessary for informed decision making and which can be compared against an accurate historical database or trusted expert opinion, best guess, or engineering estimate.

8. Model requirements review: ensure that the existing model, as described by today's specified requirements, models the appropriate problem. Thoroughly review and specify all current performance requirements and operational capabilities in a "requirements document". Ensure that these requirements/capabilities can be adequately evaluated and tested by CAA personnel. Make each performance requirement or

capability a verification/validation mission. Develop a "roadmap" document that matches every verification/validation mission to the requirements document by allocating personnel, equipment, suspense times and test standards to specific verification/validation missions. Establish a tracking system for all verification/validation missions by developing a matrix to highlight suspenses and cross-reference all verification/validation missions in the roadmap document against the model requirements document. Use the matrix to track all verification/validation missions through the verification/validation process.

9. During the verification/validation process, break the model down into manageable, tractable parts. The model is checked from the bottom up to ensure that modules and subroutines are constructed to support today's requirements. Data bases are thoroughly reviewed for accuracy to ensure that they reflect current equipment characteristics, force structure, equipment capabilities, tactical deployment, tactical employment, etc. Computer code is walked through and test runs on modules and subroutines verify changes as they are made. Verified portions of the model are then combined into larger sections for validation runs and evaluation. All validation testing is done with data bases updated to reflect current values.

10. Determine the specific verification/validation needs of models in use at CAA. Evaluate the techniques of reviewing simulation development, conceptual model assessment, software verification, operational validity, and data validity described below for applicability

to CAA models. Once the applicability of techniques are identified, consider the time and resources available at CAA to apply them. Select those techniques that support CAA's acceptable level of verification/-validation while satisfying time, budget, and other external constraints.

11. Review the simulation's development:

a. The purpose of the initial review is to identify the overall status of the simulation model. The simulation model is examined and a determination of its ability to simulate the system is assessed. It is important to identify any important characteristics or requirements of the system that are not adequately addressed in the current state of the simulation model.

b. Problem definition. The "problem" to be modeled through simulation (i.e., aircraft and ship requirements to support the sustained deployment of two heavy divisions to the Middle East) must be completely specified and determined. A well defined problem facilitates the construction of specific model requirements and outputs. Experienced personnel familiar with the intended use of the simulation examine the "problem" to ensure it is adequately specified. The simulation model's credibility rests upon its ability to aid decision making about the problem. Specific model output cannot be generated to answer vague questions posed by decision makers.

c. Model specification. The conceptual model of the problem is thoroughly reviewed to ensure that it is detailed, specific, and accurate. The conceptual specification must be complete and the model must address each specification with sufficient detail and accuracy if the simulation model is to be credible. The credibility of the

simulation model is judged on its ability to provide believable answers to system specifications and requirements.

d. Software engineering. The software is examined to determine how well it was constructed and implemented. Good software design calls for methodical implementation and specific adherence to standards.

12. Techniques for conceptual model validation:

The conceptual model is tested to determine if algorithms, assumptions, concepts and ground rules are accurate. The level of detail and fidelity in the model is assessed. Model inputs, outputs, the range of simulation, methods of engagement, threat characteristics, and the operational environment of the model are assessed. Techniques used in the assessment include subjective analysis, historical review, empirical testing, and logic traces.

a. Design credibility. The design of the model is examined to determine if it is robust, testable, affordable, effective, maintainable, and useful. The model should be capable of accepting expansion and growth if CAA considers this to be necessary. Thorough modularization and standardized, disciplined coding techniques throughout the model support these requirements.

(1) Model assumptions and ground rules. Assumptions and ground rules used in the development of model algorithms are examined for accuracy. Internal data values in the simulation model are checked for accuracy. Weapons details are examined for accuracy (weapon type, ranges, ammunition types, probabilities of kill, reaction times, target engagement times, target acquisition times, survivability, movement

rates, etc.). Battle management, command, control, and communications are examined for accurate portrayal of friendly and threat doctrine and capability.

(2) Component level of detail. All modules, subroutines, and components of the simulation model are examined to ensure that they provide the correct amount of resolution and detail. Algorithms or subroutines that interface with each other are examined to ensure that they have compatible levels of detail.

b. Concept validity. The concepts used to develop algorithms, modules, and subroutines are examined for fundamental validity and accuracy.

(1) Historical derivation. Review the historical development of the model to ensure that it was based upon sound principles and decisions. If algorithms, modules, subroutines, or sections of code were obtained from previous models (forerunners of the current model) their historical credibility and performance should be reviewed. If the model incorporates friendly or threat tactics, the validity of these tactics must be reviewed against current or changing doctrine.

(2) Logic traces. Traces are made through the model's algorithms and modules to assess adherence to the conceptual model and to the operational nature of the modeled system. Specific model entities (i.e., a tank battalion) are traced through the model to ensure that their behavior in the model adheres reasonably well to actual characteristics.

(3) Face validity. Experts familiar with the system examine the model and subjectively determine its accuracy and worth. Questions they raise should be addressed through further model evaluation.

13. Techniques for software verification:

Develop a verification/validation plan for software. Ensure that each verification/validation requirement is conducted to a specific standard. Utilize experts to conduct the code checks.

a. Examine all of the code prior to initiating verification/validation. Categorize modules, subroutines, and algorithms as to their importance in determining model reliability. Allocate verification/validation efforts and resources accordingly

b. Software development standards. Examine the underlying standards used to construct the software. If a particular military standard was used (such as Mil. Std. 2167), it is a straightforward process to walk through the code and examine it for compliance.

c. Internal software testing. Test portions or modules of code with driver programs to ensure that they function as intended. The driver program should force the code to execute through all possible decision steps and routes.

d. Correctness proofs. A mathematical proof may be constructed by an expert to verify the correctness of a module or portion of code. This technique is difficult to apply and is usually applicable only to small sections of code.

e. Automatic code checkers. Use automated code checking programs to identify uninitialized variables, inconsistent declarations, incomplete statements, infinite loops, unused sections of code, etc. Utilize automated code checking software to run traces through the code. If available, select automatic code checking software that will provide written documentation of trace results for later reference.

f. Develop and install the additional code to incorporate user intervention into the simulation model. Determine if freeze simulation, backtrack, and snapshot capabilities would enhance the simulation model's usefulness. Modify the model to generate periodic output of key parameters, variables, and conditions in hard copy form. Incorporate key parameter annunciators into the model that signal potential out-of-bounds errors in model parameters or variables during simulation. This allows the analyst to identify puzzling values and focus on particular modules or phases of the simulation where validity may be in doubt.

g. Consider utilizing redundant coding in critical portions of the model (i.e., internal self) checks to determine if parameters remain within bounds at critical portions of the simulation).

h. Ensure that the same sections of code are not reviewed several times by different analysts. Consider subroutines, functions and modules to be "black boxes." When the subroutine, module, or function is certified reliable it can be exempt from further verification/validation.

14. Techniques for operational validity:

a. Input-output relationships. Search for and examine the causal relationships between model output and internal values or inputs. Systematically determine and categorize the relationship between input parameters and output results. Determine if unrelated and independent inputs cause changes in outputs that should not occur in the "real world".

b. Event validity. Review a simulation run by experts to determine if the sequence of events generated by the model are believable and occur in a logical sequence or pattern.

c. Turing tests. Execute a simulation with carefully controlled input data then have the output reviewed by experts. If the expert can readily categorize the simulation output from real world occurrence or estimated real results then the simulation may not be credible.

d. Delphi technique. A panel of recognized experts reviews the output of the simulation model. The panel's consensus establishes the credibility of the model.

e. Demonstration technique. Execute the model to perform with a selected data base under a certain scenario for which the data base was designed. Examine how well the model performs the simulation.

(1) Simplified assumption testing. Run the model under simplified assumptions which are known to be true and examine the output for correct response. Remove modules or portions of code and examine response. Vary one input parameter or set of parameters that effect a specific assumption of the model, then examine the model output to see if the correct response was generated.

(2) Animation. Utilize graphics to examine the model's entities and events to see how they change and move through time.

(3) Predictive validation. If applicable, construct physical models of the entire system or portions of the system. Compare the performance of the physical model to simulation results under the same initialization conditions.

f. Analytic techniques. Use quantitative statistical tools (hypothesis testing, analysis of variance, analysis of covariance, multivariate techniques, regression analysis, response surface techniques, nonparametric techniques, discriminant analysis) to compare simulation results with CAA's established standards.

(1) Compare overall results to a standard. Develop standards for expected simulation response, under various initialization conditions, by collecting data from actual tests, historical records, or best estimates of the actual system. Then compare the simulation output, under each set of initialization conditions, to the standard. Define each comparison to be made between simulation output and real system performance. Generate a set of initial conditions for each comparison. Identify statistical requirements (confidence levels, significance levels, etc.) and determine the number of replications to be made. Then utilize the statistical techniques mentioned above to analyze the relationship between model output and the "real" standard.

(2) Comparison to test data. If specific test data is available for the real system, the model can be initiated to reproduce, as closely as possible, the conditions under which the test data was obtained. Model output can then be compared to specific test data by statistical methods.

(3) Sensitivity analysis. Modify input values or parameters over their entire range and examine model output. Identify those inputs that cause significant changes to output response and examine them carefully for accuracy.

(4) Predictive validation. Produce a prediction of simulation results by scientific analysis, engineering analysis, other simulation, or actual system test. Then compare simulation output to the predicted response.

g. When validating the model, conduct simulation runs with key model parameters and variables initiated over a wide range of their

permissible values to ensure that corresponding output remains within credible limits (extreme condition tests).

h. Compare simulation results to available real world data. Results may be compared to similar simulation model results, military exercises (i.e., REFORGER) and the "best estimates" of recognized experts.

i. Establish reasonable bounds for model outputs and internal parameters. Check the model to ensure that the bounds are not exceeded. If possible, build a self-check algorithm into the model to signal out-of-bounds errors.

j. Conduct degeneracy checks on key parameters and algorithms to ensure that changes in input variables lead to changes in output in the right direction (i.e., increasing the number of tanks in Blue tank battalions should result in increased kills against Red forces. If model output reflects a decrease, then troubleshooting is needed in particular subroutines or algorithms).

15. Techniques for data validity:

a. When a historical data base of system performance is not available for comparison to model output (i.e., general war in Central Europe in the 1990's), generate a "best estimate" data base of system response from similar historical events, the physical limitations and capacities of equipment, the results of similar simulations, military exercises (i.e., REFORGER), and the experience of military experts. Run a full scale simulation of the model with initial values of parameters and variables at their most likely values and compare output results to

the "best estimate" response. Consider the model to be credible if the model output closely resembles the "best estimate" response.

b. Review the military model by experts. Ensure that the simulation data base is modeled after available hard empirical data as much as possible (i.e., initial conditions for the quantity of diesel fuel in the basic load of a tank battalion should be checked against current MTOE. Subroutines governing fuel consumption should be patterned after historical usage reports experienced by tank battalions). Considerable effort should be expended to ensure that the model's initial conditions are accurately maintained in the data base.

c. History. Review the history of input data (where it was obtained and who obtained it) to access its validity for the model.

d. Internal consistency. Check all input data against a range of acceptable values. During simulation, monitor them to ensure they do not go out of bounds.

e. Accuracy of implementation. Examine input data constants for accuracy. Ensure that input random variables are of the correct probability distribution (Poisson, exponential, etc.).

f. Portrayal of constants. Ensure constant values are installed correctly, addressed properly, and remain unchanged through the simulation.

g. Justification of distributional form. Utilize statistical goodness-of-fit tests to verify the input distributions of random variables.

h. Consider and examine each specific input that goes into a military model at CAA. Ensure that the input data base accurately reflects the physical and/or engineering limitations of military

equipment, personnel, and tactics. For example, if the model includes the shipping of war materiel in existing merchant fleet vessels, ensure that cargo capacities, steaming times, load times, and unload times for each class of vessel are accurately entered into the input data base.

i. Gather historical data to construct an "output data base" for comparison to simulation model output. If a "real world" data base for system performance does not exist, use expert opinion, engineering best estimates, inference from similar historical events or systems, recent modern conflicts (Falklands, Israeli intervention in Lebanon, etc.), and military exercises (REFORGER, RED FLAG, etc.) to construct the output data base. Compare simulation results to this output data base. Based upon differences, develop prescribed values for what the model output should look like and use them to "calibrate" the model. Incorporate sensitivity analysis into this process. Modify variables, mathematical formulae, variable relationships, modules, and subroutines to "tweak" model output.

16. Documentation: Document all verification/validation activities and maintain for future review of the model.

17. Continuously request, monitor, and review user critiques of the model for their opinions of model accuracy. Analysts at the grass roots level may identify particular simulation runs that appear to be unreliable under certain initial conditions and data sets. These initialization conditions and data sets may have escaped scrutiny during initial verification/validation.

VI. Statistical Methods

A. Purpose

Appropriate output analysis of a simulation provides a specified degree of confidence on accuracy. However, there are many limitations to the use of statistical methods in large scale simulation models. The two most important limitations are the (1) high cost (in terms of time and computer usage) of performing replications of the simulation and (2) the lack of real world data from which to draw comparisons about the simulation results. Because of these limitations, it is usually impossible to perform a complete statistical analysis on a large-scale simulation.

However, there are several statistical methods that can be used to increase the understanding of the model and obtain some level of confidence that the model is correctly imitating the system being modeled. These methods focus on important modules or factors in the large-scale simulation. If credibility of these key modules can be enhanced, then the overall simulation will have increased credibility. This section focuses on six methods which have potential application: control charts, acceptance sampling, fractional factorial analysis, cluster analysis, regression analysis, and time-series analysis.

B. Control Charts

GENERAL

Control charts provide a means for determining if a particular system is statistically within control and thus predictable. "A process is described as in control when a stable system of chance causes seems to be operating" (Grant and Leavenworth, 1988). In fact, a control chart can be defined as ". . . a graphic representation of the variation in the computed statistics being produced by the process." (Wadsworth, et. al., 1986). As this broad definition implies, there are many types and uses of control charts. "Control charts provide information in three areas, all of which need to be known as a basis for action. These are (Grant and Leavenworth, 1988):

1. Basic variability of the quality characteristic
2. Consistency of performance
3. Average level of the quality characteristic"

The types of control charts that seem most applicable to large simulations are the range (R) chart used with either the average ($\overline{X}$) chart or the individual (X) chart. The R chart measures process variation and the $\overline{X}$ and the X chart measure central tendency. The R chart is usually used in combination with the $\overline{X}$ or X chart.

ASSUMPTIONS

There are not many assumptions needed to employ control charts. Generally, the data must be normally distributed; but as long as there are sufficient data points (usually only 5 or 6 in a sample which forms one data point), the Central Limit Theorem provides for a close enough approximation to normality. Methods have been developed to insure that

this normality assumption is met if individual control charts (sample size of 1) are used (Banks, 1989). A second assumption is that the runs or data points being used are independent.

Additionally, the user is required to state the magnitude of a change that should be detected in the average or the variation. For example, the control chart will have smaller bounds if a user wants to be 99% sure that a change is detected instead of only 90% sure. In order to develop these limits, a user must determine what the cost is of not detecting a change compared with the cost of needlessly examining the system.

USES

The uses of control charts in large scale simulations are many and are limited only by the imagination. There are two advantages to using control charts to analyze simulations. First, replications are not required. If a long simulation is run, where the factors to be studied are calculated frequently (or could be calculated frequently), the control chart can measure the variability and mean of the factors as the simulation progresses and the user can be confident that these factors are being developed correctly. Second, by monitoring the mean and average of important factors, understanding of the model should increase. Some sample applications are listed below.

(1) Analysis of input - This appears to be an excellent area for application. For example, the user presumably knows the maximum, minimum, and mean number of a certain weapon type that a unit of a specific size may have. If this data is being used as input, a control chart can easily highlight any change in the median or in the specified range so that appropriate action can be taken. For example, if a typical

Blue force division has x ± y tanks, and the number of tanks being input lies outside of this bound, the use of a control chart can cause an error or warning message to be printed and/or stop the program. Attribute charts also could be used (but are not discussed in this section) if a user is willing to have a percentage of the data be incorrect. Attribute charts are normally used to measure the number or proportion of items that are nonconforming. In inputting data to a simulation, the goal is to have all data conform to the standard. So, a variable control chart that stops the program as soon as a nonconforming input is detected, instead of keeping track of the number or percentage of nonconforming input, is more appropriate than an attribute control chart.

An extension to monitoring a one time input of data at the beginning of a simulation is to use control charts when data is input frequently to a program. For example, low resolution models use ATCAL input whenever a battle occurs. By monitoring each input from ATCAL, an analyst can insure that the input remains within specified bounds over time.

(2) Analysis of model processes - This method appears to be a viable use for control charts. While a program is running, key factors such as attrition, kill rate, travel rate, etc. can be monitored by a control chart. An error measure can be printed if these values exceed the control limits. For example, it is known that dismounted infantry cannot move more than x kilometers per day. If a control chart monitors the movement of a dismounted unit during the course of a simulation, and on a specific day the unit moves more than x kilometers, the amount of distance the unit moved can be printed, along with the factors that went into the movement equation. An analyst would then be able to determine if the input to the movement equation was incorrect or if the equation

itself is overestimating travel. The level of detail of the control chart is flexible. For example, a control chart can be implemented when a unit is moving up hill. By having the program measure the average distance traveled, the analyst can insure that the movement equation correctly accounts for hilly terrain. It should be pointed out that the first portion of this example relied upon the analyst knowing the maximum movement of a specific unit; while the second portion was based on the program calculating the average and range of the movement up a hill. Both types of control charts, known value and unknown value, will be discussed below.

(3) Analysis of output - All outputs with reasonable medians and variances can be monitored with discrepancies presented. One advantage in employing control charts for output is provision of the median and range. This allows a user to understand the model by observing what ranges on key factors are generated by the model and how these factors vary as compared to the outcome of the simulation.

$\overline{X}$ CHART AND RANGE CHART

The mean and range control charts are constructed under the assumption (1) that the mean and range are unknown or (2) that the mean and range are known.

Prior to developing an $\overline{X}$ chart, the variability must be in control. Variability is measured by the range chart. It is a common practice to use the centerline $\pm$ 3 standard deviations (3σ) in developing the range for control charts. If the range is unknown, the following equations describe the range chart:

$$CL_R = \overline{R}$$

$$UCL_R = \overline{R} + 3\sigma_R$$

$$LCL_R = \overline{R} - 3\sigma_R$$

where CL_R, UCL_R and LCL_R are the centerline, upper control limit, and lower control limit for the R chart, $\overline{R}$ is the average range, and σ_R is the standard deviation of the range.

These equations are generally implemented using the following:

$$CL_R = \overline{R}$$

$$UCL_R = D_4\overline{R}$$

$$LCL_R = D_3\overline{R}$$

where D_3 and D_4 are tabulated values which appear in virtually all quality control texts. (When n = 5, $D_3 = 0$ and $D_4 = 2.114$.)

If the standard deviation of the process under consideration is known, the following equations can be used:

$$CL_R = d_2\sigma$$

$$UCL_R = D_2\sigma$$

$$LCL_R = D_1\sigma$$

where d_2, D_1, and D_2 are tabulated values. (When n = 5, $d_2 = 2.326$, $D_1 = 0$ and $D_2 = 4.918$.)

If the mean is unknown, the control chart provides information as to whether the system is consistent; that is whether the mean remains within statistically specified bounds. If the mean is unknown, and the $\overline{X}$ chart is to be combined with a range chart, the formulas for the upper control limit (UCL) and lower control limit (LCL) are:

$$UCL_{\overline{X}} = \overline{\overline{X}} + 3\sigma/\sqrt{n}$$

$$CL_{\overline{X}} = \overline{\overline{X}}$$

$$LCL_{\overline{X}} = \overline{\overline{X}} - 3\sigma/\sqrt{n}$$

These equations are generally implemented using the following:

$$UCL_{\overline{X}} = \overline{\overline{X}} + A_2\overline{R}$$

$$CL_{\overline{X}} = \overline{\overline{X}}$$

$$LCL_{\overline{X}} = \overline{\overline{X}} - A_2\overline{R}$$

where $\overline{\overline{X}}$ is the average of the averages of several independent runs, R is the average range of the subgroups used to determine $\overline{X}$, and A_2 is a tabulated value. (When n = 5, A_2 = 0.577.) It is recommended that X consist of 15 to 25 samples (Wadsworth, 1986) and that the number of observations going into each $\overline{X}$ be dependent on the variable being studied. For example, if the variable being studied is unimodal, 5 observations may be sufficient for the normality assumption to be met; however, if the variable is exponential, 15 or more samples may be required. The subject of subgroup size is discussed in detail in the Issues section below.

If a mean value and a standard deviation are known (or given as standards) the applicable equations are as follows:

$$UCL_{\overline{X}} = \overline{X}_0 + A\sigma$$

$$CL_{\overline{X}} = \overline{X}_0$$

$$LCL_{\overline{X}} = \overline{X}_0 - A\sigma$$

where $\bar{X}_0$ is the known mean and A is a tabulated value. (When n = 5, A = 1.342.) This type of control chart (with values known) is used to ensure that the mean stays within desired limits.

X - CHART AND MOVING RANGE CHART

As mentioned above, when there are insufficient sample points to use 5 or so observations in one data point, a control chart for individuals may be more appropriate. This control chart uses an estimate of the standard deviation of two successive items; that is, $\hat{\sigma} = \bar{R}/d_2$. The control limits for the range chart are derived using the same equations given above for UCL_R and LCL_R. The equations for the X chart are:

$$CL_X = \bar{X}$$

$$UCL_X = \bar{X} + 3\bar{R}/d_2$$

$$LCL_X = \bar{X} - 3\bar{R}/d_2$$

where d_2 is a tabulated value, mentioned previously. If standard values are given, the control limits are (Banks, 1989):

$$CL_X = \bar{X}$$

$$UCL_X = \bar{X} + 3\sigma$$

$$LCL_X = \bar{X} - 3\sigma$$

APPLICATIONS TO ATCAL

There are many possible uses of control chart techniques in understanding and/or validating ATCAL. The use that an analyst applies depends primarily on the purpose of the ATCAL run being made. Listed below are several possibilities for using control charts in the ATCAL model attached at Appendix A (different versions of ATCAL exist). Primarily, these uses can be divided into (1) those which an analyst would use if attempting to validate ATCAL and (2) those which an analyst

can use if ATCAL is being used in a larger model to determine the outcome of a battle using high resolution input to ATCAL and a scaled output to a lower resolution model.

(1) If a user were attempting to understand and/or validate ATCAL as a stand-alone model, the first area of concern should be the input. ATCAL has a few built in checks (such as probabilities of a kill being between 0 and 1), but there are many obvious additional uses for control charts. For example, Phase I calls for the input of the number of weapons of each weapon systems to be modeled. Depending on the force to be evaluated, the analyst could develop a control chart that would print an error message if the number of tanks, light armor vehicles, helicopters, etc., exceeded the TO & E that the particular unit could be expected to have assigned. Additionally, Phase II reads a P matrix and an A matrix developed in Phase I of the model. The analyst could develop bounds on these values and insure that grossly improper values are not being used in the simulation.

A more dynamic use of control charts would be in establishing an $\overline{X}$ and range chart over a variable such as Z (in Phase II, Z = KT(k)/VNBR(k) and is used to update VNBR(k). KT is a matrix of vehicles killed and VNBR is a matrix of the number of vehicles.) This variable is selected because it is not given as output of the model, but is essential in calculations leading to the final output of the model. The X and R charts would allow the user to print the mean and variance of Z for different weapons systems, thus gaining a better understanding of how the model works and building confidence that Z remains within expected bounds. For example, problems with Z could be identified through its charting, thus providing a starting point for attempting to identify

reasons for unexpected output from ATCAL. This procedure can be extended to any of the internal variables used in ATCAL.

Virtually all of the variables that are given as output to the model can be monitored with control charts. This seems redundant because an analyst can rapidly scan the output and save the programming effort and computer time of having X and R charts made. However, if the model had massive amounts of output, so that it would be impractical to actually scan the output, control charts could be used on these parameters.

(2) If ATCAL were being used to extrapolate attritions from a high resolution model to a low resolution model, then none of the variables calculated within the ATCAL algorithm would be seen as output. In this case, a control chart on the difference of VN(i) (number of vehicles) and VNBR(i) (average number of vehicles), as well as on the other variables currently listed as output from ATCAL (Appendix A) could be used. VN(i) - VNBR(i) represents the attrition for the ith type weapon. An X control chart used in conjunction with an R control chart would allow the attrition in a particular situation to be determined so that an expert could determine if the attrition "made sense." A second way of developing the control chart for this type of situation would be for a known range of attrition, as established by benchmarks, to be programmed into the system. A standard of benchmarks could be developed from historical experience, as shown by McQuie's study (1988). An alternative, or a concurrent system, would require the development of these benchmarks by polling experts. As shown in McQuie's study, a portion of which is attached at Appendix C, the range of initial condition values for which a given benchmark is valid is large. Hence, the number of benchmarks to be developed, to cover various possible

force structures when entering a battle, is a manageable number. Again it must be emphasized that the analyst would use whichever system best met the needs of the study being performed. Without the use of control charts, the user would be accepting the fact that this complicated module, ATCAL, was performing correctly. With control charts, a one line printout of the average attrition and range of attrition over several battles of several weapon systems would allow the user to have confidence that the ATCAL portion of the program was performing as expected. Any of the other final variables developed by ATCAL, such as F (rounds fired), or Rate (rate of fire of a weapon system) could similarly be charted so that the user can gain confidence that all of the critical variables remain within expected bounds over several battles.

EXTENSIONS TO OTHER MODELS

Control charts have application to more complex models. Again, the greatest potential to be in monitoring key variables that are used in the calculation of the output of the model. This application allows the user to feel confident that the intermediate calculations are being performed correctly and to better understand the model by displaying the mean and variance of variables arising in these intermediate calculations.

The main problem in implementing control charts for these large models is the additional programming time in developing the code and the additional computer time in running the program. For these reasons, a programmer would have to be judicious about the selection of variables to be monitored under a control charting procedure.

ISSUES

Listed below are important issues of which a user must be aware. Decisions prior to implementing control charts are required.

(1) The user must determine which type of control chart to use. As discussed above, the range and mean chart should be used together. If the analyst decides not to compute $\overline{X}$ and $\overline{R}$ within the simulation, it must be decided whether to assume that values for the mean and range are known by polling experts or by using a study of benchmarks such as that conducted by McQuie. Some computer coding can be avoided if the mean and range are known. The problem with this approach is that there is no guarantee that the source of this "correct" data are in fact correct. It should be mentioned that research exists on how to combine experts' opinions. One such example is the development of a utility function by Barlow, et al. This method relies on determining a measure of the effectiveness of each experts' opinion and combining the estimates by weighting each estimate with this effectiveness factor. The user will have to determine which method best suits the purpose of the study.

(2) Stochastic simulations generally involve random variation and will generate results outside of the control limits. Hopefully, these outliers will be rare with no consistent pattern. A user will have to determine when to stop the program and analyze outliers. A control chart tells when to look for a cause of variation, but it does not tell where to look. Only the mean may change, or the variation may change, or both mean and variation may change. Additionally, both may only be out of the specified bounds once, irregularly, or in a constant pattern. A constant pattern indicates that the values being used are not independent. Additional rules that may be considered for determining

when to search for a shift in the parameters are (Grant and Leavenworth, 1988):

a. Whenever in 7 successive points on the control chart, all are on the same side of the central line.

b. Whenever in 11 successive points on the control chart, at least 10 are on the same side of the central line.

c. Whenever in 14 successive points on the control chart, at least 12 are on the same side of the central line.

d. Whenever in 17 successive points on the control chart, at least 14 are on the same side of the central line.

e. Whenever in 20 successive points on the control chart, at least 16 are on the same side of the central line.

It must be noted that different authors present different rules, but the user must be aware that only checking for values that exceed the specified ranges is insufficient. Additionally, if all of these rules are used, the chances of a false alarm increase.

The user must determine what to do when the control limits are exceeded. Ideally, when the control limits are exceeded, certain important data will be printed so that a speedy analysis can be made. If the program being used could run backwards, the user could recreate the conditions that led to the control limits being exceeded. A user may require the program to print data only if two or more points in succession exceed the control limits. This step will save computer time at the cost of less accuracy in the monitoring of the control limits. (For example, it is possible that every other value will exceed the established limits, and the program will never report this fact.) The

user will have to balance the cost savings against the specific purpose for performing the simulation (and hence the accuracy required).

(3) The user must determine the accuracy required. The equations given above are based on the limits $\mu \pm 3\sigma$. This limit is based on the fact that under normality, approximately 99.73% of the data will be within this range. However, if a user requires less accuracy, the control charts can be set at $\mu \pm 2\sigma$ or $\mu \pm 1\sigma$. This will decrease the range of allowable values, thus causing more values to be outside of the control limits, causing the user to have to make more decisions on whether the out-of-control value was random or caused by inaccuracies in the simulation.

(4) If the value for the mean is not known, subgroups must be used to determine a value for $\bar{X}$. Because there are different types of variation in most processes and because of autocorrelation in a simulation, the manner in which this data is collected is very important. "Generally speaking, subgroups should be selected in a way that makes each subgroup as homogeneous as possible and that gives the maximum opportunity for variation from one subgroup to another" (Grant and Leavenworth, 1988). Variation may be over long terms, such as weeks or months; or short-term variability over a period of hours or days. If large subgroups are taken, this short term variability may be missed, but if small subgroups are chosen, the long term variability may not be detected. If a war in Europe lasting several months is being modeled, then short term variablity in relatively quick division battles may affect the outcome of the entire war. Control charts are ideal for measuring both types of variability, but the user must design the experiment so that the variability of choice (long or short term) will

be measured. Wadsworth, et al. recommend using ". . . subgroup sizes of 4, 5, or 6 with 3 sigma control limits. . ., and with a frequency of checking based on knowledge of various changes occurring in the process . . ." One choice for determining subgroup size, is to determine the batch size that provides uncorrelated means. Several methods exist for this procedure (Law, et. al., 1982). If the individual control chart is used, no decision has to be made on determining logical subgroups, but the correlation involved in simulation can affect the results.

(5) The normality assumption must be approximately met for the variable under consideration. For example, if one were developing a control chart for uniform interarrival times, the above formulas would not apply. However, the averages of 5 or so uniform arrival times will be normally distributed. The user must be aware of this distinction, especially when using the $\bar{X}$ control chart.

(6) Lastly, just because the values remain within the bounds of a control chart does not mean that they are valid. If a control chart is based on the equations for unknown mean and range is being used, the mean and range developed by the program should be printed and verified for accuracy. It is possible that a unit is only moving x kilometers a day when not engaged, whereas we would expect the unit to move y kilometers a day. Thus, if results close to x are used throughout the program, the control chart will not indicate any error; but if x is significantly different than y; the simulation is still incorrect.

Additionally, runs also indicate out-of-control situations. Tests should be conducted for runs-up, runs down, runs above the mean, and runs below the mean to insure the process satisfies there specifications.

REFERENCES

Banks, Jerry, (1989), Principles of Quality Control, John Wiley, New York.

Barlow, R. E., Mensing, R. W., and Smiriga, N. G., (1986), "Combination of Experts' Opinions Based on Decision Theory", Reliability and Quality Control, A.P. Basu (editor), p. 9-19, Elsevier, (North Holland), New York.

Grant, Eugene L., and Leavenworth, Richard S., (1988), Statistical Quality Control, 6th ed., p. 8, McGraw-Hill.

Law, Averill M., and Kelton, David W., (1982), Simulation Modeling and Analysis, p. 295, McGraw-Hill, New York.

McQuie, Robert, (1988), "A Set of Templates for Evaluating Wargames, (Benchmarks)," U.S. Army Concepts Analysis Agency, Bethesda Maryland.

Wadsworth, Harrison M., Stephens, Kenneth S., and Godfrey, A. Blanton, (1986), Modern Methods For Quality Control and Improvement, p. 102, John Wiley, New York.

C. Acceptance Sampling

GENERAL

Two purposes of acceptance sampling are: (1) determine a course of action (accept or reject) and (2) prescribe a procedure which will give a specified risk of accepting a lot of a given quality level (Banks, 1989). This procedure can be exploited by an analyst who is faced with the task of validating an already completed model, substituting the word "program" for "lot". By developing a sampling scheme to select modules or pieces of code to analyze, the analyst can gain a measure of confidence that the entire program is accurate. The results of this sampling will allow the analyst to decide whether to reject or accept the code. Acceptance sampling capitalizes on the limited time and resources available to verify large scale models. However, the entire program will not be studied; so there is a risk that an error exists in the portion of the program not studied. Thus, acceptance sampling is an efficient means of testing a program which can reduce the work load by a large amount, while concurrently minimizing the added risk of an error.

ASSUMPTIONS

There are many acceptance sampling schemes. The attribute and continuous schemes have more applicability in verifying large scale simulations. If an attribute scheme is used, a sample of lines of code should be taken from each procedure or subroutine. Thus, each major portion of the program will be evaluated.

USES

The best way to show the power of acceptance sampling is with an example. Suppose that an analyst is faced with verifying the accuracy of

a 100,000 line program in a limited amount of time. A reasonable approach to this problem is to sample 500 lines from each module and reject the module if more than 2 errors are found. If a program has 20 modules, the amount of code that is studied is reduced by a factor of 1/10, from 100,000 to 10,000 lines.

Several items an analyst would look for when reviewing the code are:

(1) the algorithm must perform the task for which it is designed,

(2) the coding must be correct, and

(3) the section of code must be well documented.

If any of these areas do not meet the established standard, the module would be rejected. If a module is rejected, the entire model should be considered unacceptable until corrections are made.

ACCEPTANCE SAMPLING

Single-sampling

The example above is representative of a single-sampling plan. In this type of plan, a sample of code is drawn from a subroutine and inspected. If the total errors found in the sample is below a specified number, the subroutine is accepted; otherwise, it is rejected. Thus, the three parameters that must be specified are the number of lines of code in the subroutine, N. Next is the number of lines, n, to be inspected. It will be easier for an analyst to examine random blocks of code rather than n random lines because the derivation of variables can be traced. The last parameter required for acceptance sampling is the acceptance number, c, which is the maximum number of errors allowed in the sample.

In designing a single-sampling plan, an analyst would have to specify four parameters (Banks, 1989):

(1) α - the probability of rejecting a module that meets the specified quality level

(2) β - the probability of accepting a module that does not meet the specified quality level

(3) P_1 - the fraction nonconforming value for which the probability of acceptance is high

(4) P_2 - the fraction nonconforming value for which the probability of acceptance is low.

After specifying these values, n and c that will meet the conditions represented by the four parameters above can be determined.

Alternately MIL-STD-105D can be used. By specifying the lot size and inspection level, the sample size n is determined. Then, specifying the AOQL (Average Outgoing Quality Limit) leads to determination of c.

Continuous Sampling Plan (CSP)

This type of sampling is best used when there are no obvious breaks in the code or if production of the code is continuous. For example, if several programmers were working on a program and periodically turning in completed portions to an analyst, the analyst may apply a CSP to verify the code. The single-level continuous sampling procedure prescribes alternating between sequences of 100% inspection and sampling inspection (Banks, 1989). The procedure starts with 100% of the code being inspected. As soon as a prescribed number, i, of lines have been found to be correct, only a fraction, f, of the lines are inspected. When an error is found, it is corrected, and 100% inspection begins again until i lines pass inspection again.

The amount of risk in this scheme is a function of f and i. There are many combinations of f and i for the amount of risk an analyst feels

is warranted in the situation. These values are tabulated. The choice of i and f should be based on practical considerations. To study blocks of code, i should be made relatively large, so the corresponding f will be small. During fractional sampling, the lines of code to be inspected must be chosen randomly (Banks, 1989).

APPLICATIONS TO ATCAL

The version of ATCAL used in this section is made up of 1796 lines of code (excluding comments) and 15 subroutines. Two of the subroutines, Phase 1 and Phase 2 dominate the program. Both methods of acceptance sampling discussed above will be portrayed.

Suppose that an analyst wants to accept this program with no more than 1% of the lines containing errors. Using MIL-STD-105D, a solution is to let the sample size n = 125 while c = 3 if the module in question contains between 1,000-3,000 lines. Thus, each module (subroutine, function, or main program) in ATCAL would have 125 lines studied. Modules that are fewer than 125 lines would be completely studied. As long as there were 2 or fewer errors, the requirements of the analyst would be met. The number of lines of code to be studied is reduced to 80 with c = 2 if the module being inspected contains between 500-800 lines.

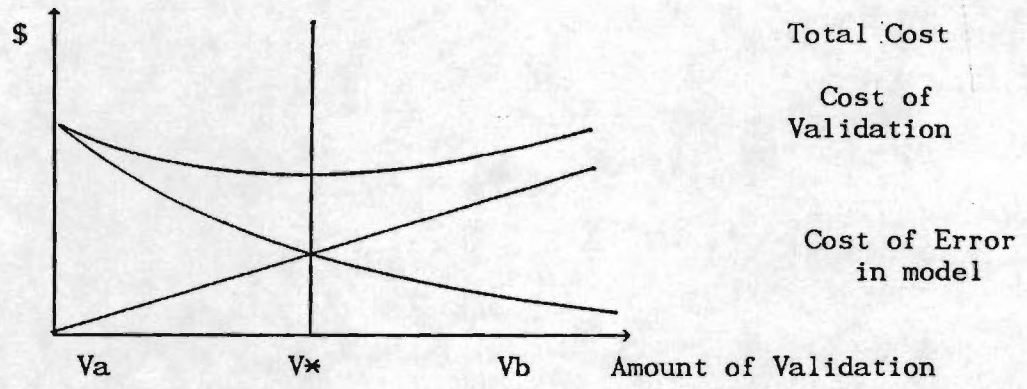
A second example using acceptance sampling is a check for the accuracy of documentation for a program. A documentation error exists if the purpose of the line of code the analyst is studying is not explained in either the internal or external documentation of the code. Also, an error exists if the documentation incorrectly states what is occurring in the program. Finally, an error exists if the documentation indicates that an outdated algorithm is being used.

EXTENSIONS TO OTHER MODELS

There should be no problem in extending this procedure to other models. As a matter of fact, this procedure is designed to save the analyst time in verifying models, so the larger the model, the more time is saved if sampling techniques are used. Because ATCAL is a relatively small program, the reduction in the lines of code that have to be analyzed is not dramatic. However, in a 100,000 line model, gigantic savings can be expected. Additionally, there are other uses of acceptance sampling plans. For example, an analyst could use a CSP if reviewing code as it is written. Suppose the analyst wants to be 99% sure there are no errors in the code and wants to study blocks of 100 lines of code. Thus, i is 100 and the frequency, f, is 1/9 (many texts contain tables for obtaining the values of f). Thus, if the first 100 lines of code were found to be error-free, only 1/9 of each remaining section would be studied. Also the analyst must verify that all algorithms in the portions of the code being studied perform correctly.

ISSUES

The main issue a user faces is accepting the risk, however small that may be, that part of the code is incorrect. The analyst will have to determine what the cost is of accepting an error in the program, and obtaining incorrect results. However, even if sufficient time and manpower are available, it still may not be cost effective to verify every line of code. As the chart below shows, there is an optimum level of work that should be performed (V*) on verifying and validating a model. This level balances the cost of an error in the simulation with the cost of verification and validation and indicates the point where the total cost is minimized.



Acceptance sampling provides a means of minimizing the portion of the cost spent on verifying the code. An analyst has the option of selectively verifying code and statistically having only a small risk of an error existing within the program, thus, freeing resources to be used on other credibility measures discussed in the verification and validation section of this paper.

REFERENCES

Banks, Jerry, (1989), Principles of Quality Control, John Wiley, New York.

D. Fractional Factorial

GENERAL

A fractional factorial design of an experiment allows an analyst to determine the significance of several factors while performing a minimal number of replications of the model being studied. Factors can be defined as the parameters and variables changed during the simulation run. The levels of a factor are the values of the corresponding input. So, a quantitative factor has many levels while a qualitative factor may have only two. In large scale models there are many factors, so a fractional factorial is preferred to a full factorial because of the number of replications that would be required in a full factorial experiment.

ASSUMPTIONS

The most significant assumption in a fractional factorial experiment is that higher order interactions are negligible. This assumption is normally valid for four or higher order interactions, but may not be true for two or three order interactions. For this reason, a fractional factorial experiment is often just a first step in analyzing the chosen factors. Identification of the minimum critical factors is an important consideration and should be based on experience of the analyst and on the purpose for which a particular model is currently being used. If the resolution of the design is chosen correctly, subsequent replications of the model can be kept to a minimum.

USES

The main use of a fractional factorial design is determining if known significant factors (real-world) are significant in the simulation.

This will allow the user to have confidence that the studied factors are determined correctly in the simulation.

2^{k-p} DESIGNS

This section relates experiments to military simulations. Their construction is left to the numerous references that discuss that subject. Fractional factorial experiments apply to both deterministic and stochastic simulations (Kleijnen, 1987).

If replications can be performed, an error term can be obtained in the ANOVA. However, if no replications can be performed, such as in a deterministic model like ATCAL, it is still possible to obtain a term that may serve as a proxy for the error term by assuming that certain factors (usually higher-order interactions) are zero. If an experiment is performed with no replications and some terms are found to be insignificant, an analyst may (but does not have to) add them to the error term, thus obtaining more degrees of freedom in the error term (Kleijnen, 1987).

A fractional factorial design is built around design generators, which are the relations used to establish the design table. It is essential that design generators be chosen so that the significance of the factors of interest can be estimated from the resulting experiment. An example of this is shown as it relates to ATCAL below. Concurrently, care must be taken to insure that the resolution of the design confounds only those factors that the user wishes to confound. As stated above, a correct selection of resolution will allow the user to focus on the factors of interest with a minimum of work if further experimentation is required. Frequently, experimentation begins with a resolution III (R-III) design. These designs apply if the factors are quantitative and

if it can be assumed that there are no interactions. The R-III design allows the analyst to determine which effects are unimportant. However, because interactions may be important, the R-III design does not allow the user to validate the significance of factors. Sequential experimentation must be performed to isolate the factors of interest. Two methods to follow after studying the results from the R-III design are:

(1) Select additional combinations so that the significance of the main effects can be determined or

(2) Increase the R-III design systematically, so that a design of higher resolution results.

The decision concerning which of the above methods to use depends on the number of factors to be analyzed and on the results of the R-III design.

If it is decided to increase the R-III design to an R-IV design, the foldover principle can be used (Kleijnen, 1987). The R-IV design results in no main effects being confounded with any other main effect or two-factor interactions. However, two factor interactions are confounded. R-IV designs yield biased estimators of the first-order effects if quantitative factors are used and if the two factor interactions are not zero. Lastly, to determine whether two factor interactions are important, an R-V design can be implemented.

APPLICATION TO ATCAL

The most obvious application of a fractional factorial design to ATCAL is to test the significance of the seven major weapon categories of the model (Appendix A). These categories are as follows : tank (Tnk), armored personnel carriers (APC), helicopters (Helo), anti-tank and mortars(AT/M), dismounted infantry (INF), artillery (ARTY), and close air

support (CAS). For simplicity, the test will be conducted with each of these weapon categories being at high (+) and low (-) levels. Next, an analyst would have to determine what output from the model to consider. It would be reasonable to monitor the attrition of the red and blue forces as the blue forces are changed from high to low levels in each of the seven categories above. One would expect statistically significant changes in the attrition rate of the red and/or blue forces as these levels are changed. Some categories, such as AT/M may only have a significant effect on the blue forces (i.e. defense) rather than on the attrition of the red forces. However, if one of these categories has no significant effect on either the attrition rate of the red or blue forces when its level is changed from high to low, the validity of the model might be questioned and further tests undertaken. It should also be emphasized that the only change in each replication is in the number of weapons in each weapon category as required by the experiment (i.e. the rate or fire of the P matrix should not be changed). Someone familiar with the system under study should be consulted to determine a realistic high value and low value (not too far apart or too close together) of each of the weapon systems to be used in the experiment, so that the model performs credibly.

A full factorial design would require $2^7 = 128$ replications. However, using a R-III design, only $2^{7-4} = 8$ replications would be required. A full factorial (k - p) design is used as the basis for a fractional factorial design. In the example given below the full factorial is a 7 - 4 = 3 design, and is found in the first three columns in the chart below. (There are numerous ways to perform a factorial analysis. The methodology that follows is from Box, Hunter & Hunter,

1978). The additional p factors to be estimated are then derived as multiples of the basic table, hence the name "design generator." One such design is given below where 4 = 12 (the ith element in column 4 is obtained by multiplying the ith element in column 1 by the ith element in column two), 5 = 13, 6 = 23, and 7 = 123 are used as generators. In the example, there are 4 additional factors, and only 4 combinations of the three factors in the basic table, so this is a saturated R-III design.

	Tnk	APC	Helo	AT/M	INF	ARTY	CAS	Attrition Blue	Attrition Red
run	1	2	3	4 12	5 13	6 23	7 123	y_1	y_2
1	-	-	-	+	+	+	-		
2	+	-	-	-	-	+	+		
3	-	+	-	-	+	-	+		
4	+	+	-	+	-	-	-		
5	-	-	+	+	-	-	+		
6	+	-	+	-	+	-	-		
7	-	+	+	-	-	+	-		
8	+	+	+	+	+	+	+		

With this design, the following effects are confounded (assuming 3 or higher order interactions are negligible):

$l_1 = 1 + 24 + 35 + 67$

$l_2 = 2 + 14 + 36 + 57$

$l_3 = 3 + 15 + 26 + 47$

$l_4 = 4 + 12 + 56 + 37$

$l_5 = 5 + 13 + 46 + 27$

$l_6 = 6 + 23 + 45 + 17$

$l_7 = 7 + 34 + 25 + 16.$

l_i represents the sum of the effects of the terms following the "=" sign. These interactions are obtained by multiplying each of the defining relations by the factor to be determined. The first step in this process is to obtain the generating relations. For example, the generator 4 = 12 can be rewritten as I = 124 where the identity I indicates all '+' signs in a column and indicates that any effect times itself is equal to I. That is 4 x 4 = I. Thus, the generating relations are I, 124, 135, 236, and 1237. Now, to obtain the defining relation, all multiples of the generating relations must be made. Multiplying all possible combinations of the generating relations, two at a time, yields I = 2345 = 1346 = 347 = 1256 = 257 = 167; three at a time yields I = 456 = 1457 = 2467 = 3567 and four at a time gives I = 1234567. This set, plus the generating relations is defined as the "defining relations." So, to obtain the effects confounded with effect 1 by: 1 x I = I, 1 x 124 = 1124 = 24, 1 x 135 = 1135 = 35 and 1 x 167 = 67, ignoring all three-factor or higher-order interactions. All other confounding relations can be obtained by multiplying them by the defining relations given above.

After analyzing the data (y_1 and y_2) to determine each of the effects, an analyst could logically determine what the next step of analysis should be. The key point at this stage of the analysis is to proceed in a sequential process, according to one of the two methods described above to systematically reduce the number of factors being considered, using a minimal number of replications at each step in the process. It must be remembered that due to confounding, at this point an analyst cannot determine the significance of any of the main effects. Thus, if the effect of l_1 appeared significant, an analyst would not know

if this result were due to tanks, or to the interaction of APC's with anti-tank/mortars, or the other interactions listed on the first line above. However, if an effect were insignificant, the analyst could judge that the particular main effect and all of the interactions confounded with it were not important to the model.

Assuming that the analyst was interested in the main effects, a R-IV design could be created by changing all of the signs (foldover) in the table listed above. The confounding pattern would then be:

$l'_1 = 1 - 24 - 35 - 67$

$l'_2 = 2 - 14 - 36 - 57$

$l'_3 = 3 - 15 - 26 - 47$

$l'_4 = 4 - 12 - 56 - 37$

$l'_5 = 5 - 13 - 46 - 27$

$l'_6 = 6 - 23 - 45 - 17$

$l'_7 = 7 - 34 - 25 - 16.$

Thus to determine the main effect of tanks, one could simply apply the equation of $1/2(l_1 + l'_1)$, and by changing the subscript, obtain all of the main effects. So, in just 16 runs, all of the main effects can be obtained and their significance judged. If just one column had its sign changed in the table above, an analyst could determine the main effect and all of the two-way interactions associated with that one factor of interest.

Again, this analysis would allow an analyst to fail to reject the given model as being invalid, and would indicate that the model was handling important factors in an expected manner. It would also increase understanding of the model because an analyst could point out that some factors in the model were insignificant. A decision could then be made

on whether to include those factors in the model, perhaps simplifying code and/or the model.

EXTENSIONS TO OTHER MODELS

The major problems with extending this method to other models is the number of replications required and the large number of factors involved. Specifically, if a model takes several days to run, then even 16 replications may be impractical. Similarly, with a large number of factors involved, more runs will be needed to screen out the unimportant factors, and to determine the degree of interaction with factors not considered in the experiment. In the "ISSUES" section below, a means of using a fractional factorial design to screen a large number of factors is given.

There are at least two courses of action to alleviate these problems. Known or expected values can be substituted for some factors. If this procedure is implemented, computer time can be saved because random numbers will not have to be generated. The factors replaced by expected values will not impact on the fractional factorial design, so the number of factors to be considered will be reduced. Second, it may be possible, as was explained above, to perform factorial experiments on modules of the large model. Knowledge of the functioning of key modules will allow an analyst to express increased confidence (or lack of confidence) in the entire model (Haley and Ghelber, 1980).

ISSUES

(1) Probably the most critical issue involved is the huge number of factors involved in large scale simulations. One solution to this issue is to use group-screening designs (Kleijnen, 1987). In this design, individual factors are aggregated into groups of factors. If a group is

not significant, it is concluded that all of its factors are unimportant. A group is considered to be at its low level if all of its members are at their low level and is considered to be at its high level if all of its member are at their high level.

In a group-screening design, it is assumed that the signs of the main effects are known. That is, it can be insured that all factors in a group are at their high or low level. Further, it is assumed that the interactions among the effects in a group are unimportant. Therefore, the effects of the individual factors within a group cannot compensate for one another. In order for this assumption to not effect the results, a R-IV or higher design should be implemented when analyzing the groups.

Once it has been determined that certain effects are unimportant, the factors in that group should be kept at a fixed level throughout further experimentation. This will aid in the determination of the importance of any of the other effects. Group screening can be continued until a manageable number of factors remain for either fractional factorial of full factorial experiments. The following formula can be used to determine the optimal size of the groups:

$$f = [(1 - \alpha)p]^{-1/2}$$

where p equals the likely number of important factors divided by the total number of factors.

REFERENCES

Box, George E.P., Hunter, William G., and Hunter, J. Stuart, (1978), Statistics for Experimenters An Introduction to Design, Data Analysis, and Model Building, p. 381, John Wiley, New York.

Haley, Charles A. and Ghelber, Craig S., (1980), A Methodology For Validation of Complex Multi-variable Military Computerized Models, p. 26, Air Force Institute of Technology

Kleijnen, Jack P.C., (1987), Statistical Tools for Simulation Practitioners, p. 259, Marcel Dekker, New York.

E. Cluster Analysis

GENERAL

Cluster analysis is a means of reducing a large number of data elements into a form that highlights the relationships between the data elements. There are many algorithms that perform cluster analysis, but essentially each divides the data into groups such that the differences in the data elements within a group is less than the differences in data elements between groups. Hence, the data elements within a group can essentially be considered the same; thus reducing a large data file into a lesser number of clusters.

ASSUMPTIONS

There are no assumptions in the cluster analysis algorithm discussed below. However, a user must realize that even though a cluster of data is different from another cluster the data within a cluster is not the same. The use of the clusters must recognize this fact. The user should select an algorithm that relates the data in a way that matches the purposes of the study.

USES

Cluster analysis can be used in understanding a model by showing the relationship between input and output. A cluster analysis algorithm can be applied to several outputs from a simulation run with different inputs. The user can study the input that led to each cluster, thus determining if it is reasonable that different inputs led to essentially the same output. This analysis allows a user to understand how the model responds to different levels of input. Additionally, increased

credibility of the model can be obtained if the clusters correspond to experts' beliefs that the input should have led to similar results from the model. A Turing test could be used in substantiating the credibility of the model. Experts can be given different inputs and asked to recognize the corresponding output. If the experts' selections are clustered in a manner similar to that obtained from cluster analysis the credibility of the model would be enhanced.

CLUSTER ANALYSIS

The discussion of a basic method for conducting cluster analysis follows (Dillan and Goldstein, 1984). It is assumed that the different outputs can be expressed as numerical vectors. For example, the output from ATCAL can be expressed as $X = (x_1, x_2, x_3, \ldots x_n)$ where x_i represents the number of ith weapons system attrited, where there are n different weapon systems. The first requirement is to determine a measure of the differences between X's. One such measure is the Euclidean distance, that is $d_{ij} = \left[\Sigma(X_{ij} - X_{jk})^2\right]^{1/2}$. Next, a method is required to determine which data points, based on the distances between them, should belong to a cluster. The single linkage method uses a minimum-distance rule that starts with the two data points having the shortest distance. Call these points cluster A. Next, the minimum distance between two data elements, excluding the distance between the two elements in cluster A is determined. If this distance is between a data element in A and one not in A, X_0, then X_0 is added to cluster A. Otherwise, cluster B is begun, consisting of X_0 and the point closest to it. This process continues until all data elements belong to a cluster.

APPLICATIONS TO ATCAL

The method described above can be directly applied to ATCAL. For example, suppose 5 runs are made resulting in attrition vectors X_1, X_2, . . . X_5. Suppose the Euclidean distances are computed and result in:

	1	2	3	4	5
1	0.0	1.0	5.0	6.0	8.0
2	1.0	0.0	3.0	8.0	7.0
3	5.0	3.0	0.0	4.0	6.0
4	6.0	8.0	4.0	0.0	2.0
5	8.0	7.0	6.0	2.0	0.0

The closest elements are elements 1 and 2, so they initially make up cluster A. Distances are recomputed and result in:

	12	3	4	5
12	0.0	3.0	6.0	7.0
3	1.0	0.0	4.0	6.0
4	6.0	4.0	0.0	2.0
5	7.0	6.0	2.0	0.0

The two data elements closest now are elements 4 and 5, so they become cluster B. Distances are recomputed, resulting in the following matrix:

	12	3	45
12	0.0	3.0	6.0
3	3.0	0.0	4.0
45	6.0	4.0	0.0

So, element 3 should belong to cluster A. This completes the cluster analysis process.

EXTENSIONS TO OTHER MODELS

The most difficult decision in using cluster analysis on a model with many factors is determining which factors should be used in the clustering algorithm. Even in a model such as ATCAL, the vectors being studied could have been based on the number of rounds each weapon system fired. As the number of factors increases, the validity of the model may depend on an increasing number of factors. These important factors must be identified so that cluster analysis correctly corresponds to the output of the model.

ISSUES

Two issues involving cluster analysis have already been discussed. First, the important output factors must be identified so that the cluster analysis is valid. Second, the use of cluster analysis must be carefully monitored. For example, the difference in items discussed above is based on the Euclidean distance between data elements. However, for similarity type data (0, 1), it is more appropriate to use a matching-type measure (Dillan, 1984). The correct distance measure and algorithm must be used, based on the data elements being studied.

It should also be pointed out that most computer statistical packages, such as SPSS and SAS perform cluster analysis. So, the amount of work, other than replications of the model is minimal.

REFERENCE

Dillon, William R. and Goldstein, Matthew, (1984), Multivariate Analysis Methods and Applications, p. 161, John Wiley, New York.

F. Regression Analysis

GENERAL

Regression analysis examines the effects that certain "independent" variables exert (or appear to exert) on a "dependent" variable. Often, a functional relationship exists between these variables; whether this relationship can be described in manageable terms is another matter. Even when no relationship is present, it may be possible to relate the variables using mathematical equations. Thus, the goal of any regression analysis is to obtain a vehicle for predicting the value of a response from knowledge of those variables contributing to its outcome. For example, the equation $Y = B_0 + B_1 X_1 + B_2 X_1X_2 + B_3X_2^2 + e$ could signify the relationship between Y_1 and X_1& X_2

ASSSUMPTIONS

An assumption common to most regression analyses is that the independent (predictor) variables are not subject to random variation, but that the dependent (response) is. While this is seldom the case, it is usually true ". . . that random variation in any of the predictor variables is so small compared with the range of that predictor variable observed that we can effectively ignore the random variation" (Draper and Smith, 1981).

A second assumption frequently made concerns the distribution of the residuals (e_i) where $e_i = Y_i - \hat{Y}_i$, the difference between the response variable ($\hat{Y}_i$) and the "fitted" value ($\hat{Y}_i$). While an assumption that these errors are independent and normally distributed is not required to obtain a regression equation, it is necessary when tests are conducted to measure the equation's lack of fit. Finally, irrespective of

distribution properties of the errors, it is assumed that the coefficients of the predictor variables (B_i) are linear functions of the response variable (Y) and ". . . provide unbiased estimates (of these coefficients)" which have minimum variances (Draper and Smith, 1981).

USES

Regression analysis can be used with large scale simulations by quantifying the relationship between input and output. Given one or more variables, a regression equation can be obtained from the results of a simulation run. The user can then determine if a viable relationship exists between the different inputs and the output. Also, the impact of each variable can be measured; some may play little or no role, while others may be responsible for the bulk of a simulation's product.

This relationship, if it exists, can be used most efficiently to streamline verification efforts by providing a range of values for predictive purposes. In other words, given a range of realistic input values, a corresponding window of acceptable output values can be obtained. The more frequently a model's results compare favorably to these predicted ranges, the greater its credibility.

THE "BEST" REGRESSION EQUATION

Given that we want to establish a regression equation for a specified response (Y) in terms of the predictor variables (X_i), two opposing criteria for selecting an equation complicate matters. On the one hand, to make an equation as predictive as possible, numerous functions (X_1^2, X_1X_2, ℓnX_2, etc.) of the X_i should be considered. On the other hand, the simpler the equation, the more utility it possesses for analysis purposes.

It is for these two reasons that the stepwise regression procedure usually produces a "best" equation. This approach is more economical of computer facilities than other methods (Multiple, Backward Elimination, Ridge, etc.) and avoids working with more X's than are necessary (while improving the equation at each iteration). A simplified explanation of the procedure is as follows:

1. Calculate/compare the correlations of all predictor variables not in the equation with the response. Choose the most highly correlated variable to enter the equation.

2. Test the new regression equation for overall significance using a t or F-test. Retain the new variable in the equation if it is significant.

3. Conduct a partial t or F-test on each variable in the current equation. This will determine the contribution of each. If a variable fails to be significant, remove it from the equation.

4. Continue steps 1-3 until all predictor variables have been subjected to either removal or rejection.

APPLICATIONS WITH TRANSMO

The concept of regression analysis can be applied to a test run of the U. S. Army Transportation Model (TRANSMO). For example, consider those predictor variables which may affect the percentage of a cargo shipment ultimately reaching its destination. The basic independent variables could be:

X_1: Quantity of an asset used for transport (# of planes or ships).

X_2: Amount of transport time required (days).

X_3: Distance between point of departure/arrival (thousands of miles).

X_4: Attrition factor.

X_5: Type of asset used for transport (plane, ship or rail).

X_6: Amount of cargo shipped from point of departure (short tons).

Then, using the stepwise procedure with a regression program (MINITAB), 85 observations of the response variable were examined in terms of 43 predictor variables (X_i, X_iX_j, X_iX_j,X_k, X_i^2, $\ln X_i$)

STEP	1	2	3
CONSTANT	33.195	-7.039	50.032
X_6	0.872	0.812	0.819
T-RATIO	39.84	30.28	31.09
LNX_1		68	57
T-RATIO		3.51	2.89
X_4			-579
T-RATIO			2.28
STD DEV	164	154	150
R-SQ	95.03	95.68	95.94

The resulting fitted equation is $\hat{Y} = 57\ell nX_1 - 579X_4 + 0.819X_6$. The R-SQ term measures the proportion of total variation about the mean (Y) explained by the regression. The closer this value is to 100 (percent), the better the fitted equation explains the variation in the data. The results of this example indicate a heavy reliance on but three predictor variables.

EXTENSIONS TO OTHER MODELS

The biggest hurdle in using regression on a model with many factors is determining whether a variable is dependent or independent. Often a variable will act as both in different phases of a simulation. ATCAL, for instance, considers the level of attrition inflicted on a force by dismounted infantry. However, the effectiveness of the infantry considered is in turn dependent on such factors as movement speed and ammunition resupply.

ISSUES

Since R^2 is often used as a measure of the "accuracy" of the regression equation (where variation of the data is concerned), one must closely monitor the change in this statistic with the addition of each predictor variable to the equation. R^2 will always increase as the number of parameters in the equation approaches saturation) that is, the number of distinct observations of the response.

Secondly, the value of the correlation between the response and a predictor variable only indicates the extent to which X and Y are linearly associated. "It does not by itself imply that any sort of causal relationship exists between X and Y" (Draper and Smith, 1981).

REFERENCES

Draper, N. and Smith, H., (1981). Applied Regression Analysis, John Wiley, New York.

G. Time Series Analysis

GENERAL

Time series analysis is a means of forecasting the demand levels of a group of related items. Any forecast is based on a combination of historical data, mathematical model(s), human judgment and associated error. The focus here will be on the mathematical model and the part it can play in system verification/validation. Specifically, time series analysis can be used to determine how accurately a simulation estimates the probable range of system demand around the expected value; that is, how closely the simulation tracked historical or projected data. The mathematical model proposed for analyzing x_t, the demand in period t is:

$$x_t = (A + bt)F_t + e_t$$

where a = level of response

b = linear trend

and e_t = irregular random fluctuations

Due to the presence of random error, the exact values of the parameters can not be found; for prediction purposes this is not a concern. Consequently, the emphasis will be on estimating these values in order to verify the behavior of a system.

ASSUMPTIONS

There are two assumptions associated with time series analysis. The first is that the random errors (e_t) are independent variables with mean 0 and constant variance. It is also imperative that the mathematical model selected be appropriate for the system under scrutiny. For instance, a cyclical trend may not be present in the data; a model with less parameters would therefore be more accurate and less complicated.

USES

Time series analysis can be used in a manner similar to regression analysis. Given a set of historical "demands" or input values, a mathematical model can be generated to create a window of legitimate entries. The user can then determine if different levels of input are valid and also investigate if the simulation responds accordingly. The value of a time series analysis lies in its preventive nature; the more realistic the input the more credible the model's output.

TIME SERIES ANALYSIS

As stated earlier, with trend and/or cyclical factors present in the data, time series analysis is more complicated. This is because the effects of each must be isolated. The procedure for this is as follows:

1. Initial estimation of level/trend at each period. The trend point for a particular period t is most commonly estimated using the ratio to moving average procedure. The moving average is observed relative to a complete cycle of P periods. (A complete cycle is used in order to remove cyclical effects from the trend.) For example, in Figure 1, P = 4 and the entries in column 3 reflect the total demand for each consecutive cycle. Because the 4-period moving average ends up being centered between two periods, the average of each consecutive two moving averages is taken to further define the trend in column 4. Notice that the sixteen historical periods are ordered from 0 to 15; forecasted demand will then begin with period +1.

2. Estimate of cyclical factors. The estimate of the cyclical demand (column 5) for each period t is obtained by dividing demand (x_t) by the centered moving average. Due to the nature of the

moving average procedure in step 1, estimates of the cyclical factor can not be calculated for periods -15, -14, -1 and 0.

3. Normalization of cyclical factors. The estimates obtained in step 2 still contain random components. In an attempt to lessen their contribution, the seasonal factors are averaged for similar periods in each cycle. Because the total of these averages may not exactly sum to P, it is desirable to normalize each index as shown in Figure 2 and reflected in column 6 of Figure 1 (Silver and Peterson, 1985).

4. Estimating $\hat{a}_0$ and $\hat{b}_0$. Using the normalized indices found in step 3, the data is stripped of cyclical factors (x_t/F_t) as shown in column 7 of Figure 1. The level estimates are then fit to a regression line using the equations (Silver and Peterson, 1985):

$$\hat{a}_0 = [6/n(n+1)] \sum_t t x_t + [2(2n)1)/n(n+1)] \sum_t x_t$$

$$\hat{b}_0 = [12/n(n^2-1)] \sum_t t x_t + [6/n(n+1)] \sum_t x_t$$

For this example, $\hat{a}_0 = 76.91$ and $\hat{b}_0 = 1.68$. The model estimate is therefore $\hat{x}_t = (76.91 + 1.68t)\hat{F}_t$.

5. Forecasting The forecast of demand in any future period is obtained by the following equation:

$$\hat{x}_{t,t+\Delta} = (\hat{a}_t + \hat{b}_t\Delta)\hat{F}_t$$

Using the values obtained in steps 1)4, the forecast for the next four periods is:

$$\hat{x}_{0,1} = (76.91 + 1.68)(0.86) = 67.6 \text{ units}$$

$$\hat{x}_{0,2} = (76.91 + 3.36)(1.07) = 83.9 \text{ units}$$

$$\hat{x}_{0,3} = (76.91 + 5.04)(1.32) = 108.2 \text{ units}$$

$$\hat{x}_{0,4} = (76.91 + 6.72)(0.75) = 62.7 \text{ units}$$

PERIOD t (1)	DEMAND x_t (2)	4-PERIOD TOTAL (3)	CENTERED AVERAGE (4)	$\hat{F}_t$ ESTIMATE (5)	$\hat{F}_t$ NORMALIZED (6)	LEVEL ESTIMATE (7)
-15	43				0.86	50.0
-14	57				1.07	53.3
		217				
-13	71		55.1	1.29	1.32	53.8
		224				
-12	46		56.5	0.81	0.75	61.3
		228				
-11	50		58.8	0.85	0.86	58.1
		242				
-10	61		60.6	1.01	1.07	57.0
		243				
- 9	85		62.3	1.36	1.32	64.4
		255				
- 8	47		65.4	0.72	0.75	62.7
		268				
- 7	62		67.1	0.92	0.86	72.1
		269				
- 6	74		67.4	1.10	1.07	69.2
		270				
- 5	86		66.9	1.29	1.32	65.2
		265				
- 4	48		66.8	0.72	0.75	64.0
		269				
- 3	57		69.3	0.82	0.86	66.3
		285				
- 2	78		72.9	1.07	1.07	72.9
		298				
- 1	102				1.32	77.3
0	61				0.75	81.3

Figure 1

PERIOD t (1)	$\hat{F}_t$ SUM (2)	AVERAGE ESTIMATE (3)	NORMALZIED INDEX (F_t) (4)
-15/-11/-7/-3	2.59	0.86	0.86
-14/-10/-6/-2	3.18	1.06	1.07
-13/-9/-5/-1	3.94	1.31	1.32
-12/-8/-4/0	2.25	0.75	0.75
		----	----
TOTAL		3.98	4.00

Figure 2

6. Revision of level, trend and cyclical factors. Once actual demand for a period is realized, the level, trend and cyclical factors can be updated to reflect this additional knowledge using the following relationships (Silver and Peterson, 1985):

$$a_t = (c_a)\ (x_t/F_{t-p}) + (1-c_a)\ (a_{t-1} + b_{t-1})$$
$$\hat{b}_t = (c_b)\ (\hat{a}_t - \hat{a}_{t-1}) + (1-c_b)\ (\hat{b}_{t-1})$$
$$\hat{F}_t = (c_f)\ (x_t/\hat{a}_t) + (1-c_f)\ (\hat{F}_{t-p})$$

where the constants c_a, c_b and c_f are chosen using the following guidelines:

	c_a	c_b	c_f
Upper Limit	0.51	0.176	0.50
Reasonable Value	0.19	0.053	0.10
Lower Limit	0.02	0.005	0.05

For stability purposes, the value of c_b should be kept below that of c_a. Now, suppose that the actual demand in period 1 was 75units. In step 5, a forecast of 67.6 units was made. The revised value of the level, trend and cyclical factors would then be

$$\hat{a}_1 = (0.2)(75 / 0.86) + (0.8)(76.91 + 1.68) = 80.31$$
$$\hat{b}_1 = (0.1)(80.31\ )\ 76.91) + (0.9)(1.68) = 1.85$$
$$\hat{F}_1 = (0.3)(75 / 80.31) + (0.7)(0.86) = 0.88$$

where c_a, c_b and c_f are set at 0.2, 0.1 and 0.3 respectively.

APPLICATIONS TO TRANSMO

The method of time series analysis can be applied to TRANSMO. For instance, it may be desirable to determine if demands for ammunition/supplies by a theater are following historical or projected

levels. Using the procedure described above, an estimate of the mathematical model with level, trend and cyclical factors is obtained. The estimate (based on 40 days of historical data) takes the form $\hat{x}_t = (733.35 + 17.54t)\hat{F}_t$ as shown in Figure 3.

Based on this mathematical model, an estimate (forecast) of demand for day 1 should be in the neighborhood of:

$$\hat{x}_{0,1} = (733.35 + 17.54)(1.15) = 863.52 \text{ short tons}$$

However, suppose a more realistic figure is 1106 units. It would then be feasible to update the level, trend and cyclical constants accordingly.

$$\hat{a}_1 = (0.2)(1106 / 1.15) + (0.8)(733.35 + 17.54) = 793.06$$

$$\hat{b}_1 = (0.1)(793.06 - 733.35) + (0.9)(17.54) = 21.76$$

$$\hat{F}_1 = (0.3)(1106 / 793.06) + (0.7)(1.15) = 1.22$$

The forecast for day 2 would then be adjusted to reflect the increased level of demand:

$$\hat{x}_{0,2} = (793.06 + 21.76)(1.22) = 996.83 \text{ short tons}$$

Continuation of this process would yield a range of legitimate values for the modeler to analyze input.

EXTENSIONS TO OTHER MODELS

Time series analysis can be utilized in the verification of a model (or portion thereof) that depends on historical or projected input as a function of time. ATCAL, for instance, could be analyzed for realistic expenditures of artillery or tank ammunition based on the resupply data. Obviously, while a time series will not serve as a judge of any model's credibility, it can aid in the elimination of results due to erroneous input.

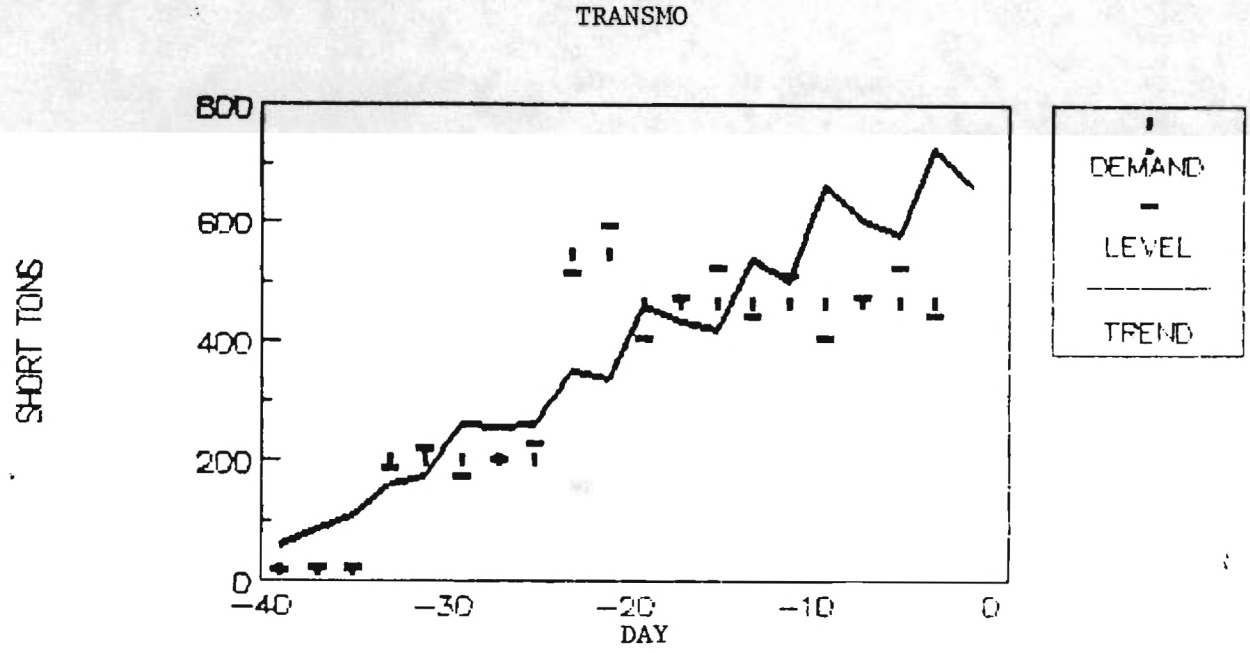

TRANSMO
SHORT TONS
800
600
400
200
0
-40
-30
-20
-10
0
DAY
DEMAND
LEVEL
TREND

Figure 3

ISSUES

The updating, forecasting and initialization of a time series model is clearly dependent on the presence of level, trend and/or cyclical factors. Analysis is difficult at best when an improper model is chosen. For example, if during the initialization stage, the cyclical factors " . . . are all close to unity or fluctuate wildly, then a cyclical model is probably inappropriate" (Silver and Peterson, 1985). Thus, the decision to use a particular model should be made carefully and only after the appropriate factors are known to exist.

REFERENCE

Silver, Edward A. and Peterson, Rein, (1985), Decision Systems for Inventory Management and Production Planning, p. 112, John Wiley, New York.

VII. Extension of Known Verification and Validation Techniques

A. Purpose

The purpose of this section is to examine several established techniques for verifying and validating computer simulations and consider their application to the large-scale military simulation models commonly developed and used at the U. S. Army's Concepts Analysis Agency (CAA). Techniques currently being used by CAA and other Army agencies are identified, how they are being implemented is reviewed and possible extensions or modifications to increase the effectiveness of these methods are discussed. Their applicability to a structured verification and validation process for CAA is considered.

B. Introduction

Use of the terms verification and validation has been prevalent in computer simulation literature since the mid-1960's. The following definitions are generally accepted as standard. Verification is considered to be the process of determining whether a simulation model performs as intended or designed; in other words, debugging of the computer program. Validation is testing and determining whether the results generated by the simulation model accurately represent the real-world system that is being studied (Law, Kelton, 1980). All of the Army agencies contacted during this research effort used the terms in the context described above.

As the use of large and complex computer simulations has increased, other relevant factors which affect the overall performance of simulation models but which are not considered to specifically fall into the category of either verification or validation have been identified. Examining or testing these factors also helps evaluate a model. Many

authors have referred to this process of verification, validation and other testing as assessing the overall credibility of the model. Part of establishing model credibility is conveying the results of testing and the significance of its results to the decision makers or managers involved in a simulation study. Model credibility is established by verification, validation and other testing if a required level of correspondence between the model and the real system being modeled is achieved consistent with the domain of intended application and the study objectives (Schlesinger, 1979). This definition of model credibility will be used throughout this section.

C. The Concept of Model Credibility

Several documents have been published recently which review the field of verification and validation of simulation models. They provide a starting point for any investigation of this topic. In March 1986, Banks et al. compiled a literature review of model credibility evaluation topics for CAA. A detailed discussion of verification and validation and their role in the modeling process and a complete listing of references was included. In a 1986 paper titled "Credibility Assessment of Simulation Results: The State of the Art" , Balci summarized and surveyed research conducted over the past 20 years in the field of computer simulation model credibility and discussed the most promising future research directions. Included in the paper is a glossary of common terminology and a listing of all the important references published since 1961 related to simulation credibility. They are repeated in Appendix D of this section. Balci discusses a "hierarchy of the credibility assessment" and defines eleven credibility assessment stages (CAS) which are critical indicators of a model's credibility. Two

of these eleven indicators are model verification and validation. The others are as follows:

1. Model qualification
2. Communicative model verification
3. Experiment design verification
4. Data validation
5. Formulated problem verification
6. Feasibility assessment
7. System and objectives definition verification
8. Quality assurance of model
9. Credibility of simulation results.

In discussing model validation, Balci provides two tables listing papers which have presented statistical validation techniques and subjective validation techniques. These tables are included in Appendix D of this paper. Balci indicates that only models of "completely observable" real systems can be validated using the more reliable statistical techniques. Unfortunately, most military simulations are models of systems which are not observable. Thus, validation efforts have traditionally been restricted to subjective methods. Balci concludes that most of the research in the past has focused on validation methods to the detriment of furthering the understanding of the importance to model credibility of the other 10 indicators. He calls for increased research in the nine areas listed above to develop more powerful methodologies for assessing model credibility.

In September of 1986, Banks et al. proposed a scoring model methodology to CAA for the verification and validation of simulation models. This methodology also emphasized evaluating the overall credibility of a model, accomplished primarily through applying a structured series of subjective verification and validation techniques throughout the development (life-cycle) of the model.

In December of 1987, the United States General Accounting Office (GAO) released a report on Department of Defense simulation models. This report reviewed several large models used to provide data about weapon system effectiveness and efforts of the DOD to ensure that these models are credible. Not coincidentally, some of the GAO's conclusions and recommendations were very similar to ideas in Balci's and Banks' work. For example, GAO used fourteen factors to evaluate overall model credibility. These factors were divided into three areas of concern and, as Balci suggested, verification and validation were listed as individual factors. Examining the GAO's framework shown in Table VI-1, it is clear that it was devised with combat simulation models specifically in mind and is very appropriate to this study.

However, in contrast to Balci, GAO stated that "a major threat to credibility (of models) is the limited evidence of efforts to validate simulation results by comparing them with operational tests, historical data, and other models." In other words, they recommended attention to all the above areas but in particular to model validation.

Another procedure for measuring model credibility was described in a 1983 paper by three faculty members of Laval University in Quebec. Factors impacting on a model's credibility were organized into five general areas of validity and could be evaluated at one of five levels of

Exhibit VI-1. GAO Framework

Areas of Concern	Factor
Theory, model design and input data	1. Agreement between theoretical approach and real events
	2. Choice of effectiveness measures
	3. Portrayal of combat environment
	4. Representation of weapon operational characteristics
	5. Depiction of broad-scale battle environment
	6. Mathematical/logical representation
	7. Input data
Correspondence of real world and model	8. Verification
	9. Statistical quality of results
	10. Sensitivity testing
	11. Validation
Management issues	12. Organizational support
	13. Documentation
	14. Full disclosure of results

satisfaction. A simple 5 X 5 matrix was used to display this overall rating of a model. The key difference between this model credibility assessing procedure and others was the establishment of desired "minimal confidence limits" for each of the five validity areas by the future user and the modelers prior to the building of the model (Landry, Oral, 1983). Thus, once each of the five general areas reaches the prescribed level of satisfaction the model is credible and should be able to accomplish its purpose. In the past, testing and understanding a computer simulation has involved using methods considered validation and verification activities. The thrust of much recent work has emphasized that other factors affect the results provided by the model and that techniques must be developed to evaluate a model's performance in these areas as well as in traditional validation and verification areas. Army agencies currently use a variety of validation and verification methods some of which are designed to measure model performance in an area related to one of the additional factors discused above. In summary, it is clear that there are areas other than verification and validation which affect model credibility and that, in general, these areas must be evaluated or monitored beginning with the initial phase of the modeling process.

D. Verification

Verification has been the least troublesome of the factors affecting complex simulation models. Determining how a model is functioning and comparing this to the original model conceptualization is possible even with very complex simulations. However, an early, systematic and modular approach is necessary if verification is to be completed in a timely manner. The most commonly accepted verification techniques are:

1. Documentation of code
2. Structured programming
3. Structured walk through
4. Traces of key model events
5. Observing known values as output
6. Operational graphics (Banks and Carson, 1984)

Most Army analysis agencies are familiar with all the above methods. In a recent article in Phalanx, the bulletin of military operations research, an approach to verification (and validation) being used at the TRADOC Analysis Command - Ft. Leavenworth (TRAC-FLVN) is outlined Flanagan, 1988). In addition to several of the above methods, a planning or review phase before model development begins is used "as a technical review to ensure suitability of the algorithms" (Flanagan, 1988). Model design, data structures and algorithm selection are all reviewed at this time. The structured walk through is described as a joint effort of the designer, coder and reviewer. An input data review is also performed. Data quality was considered a factor affecting model credibility by both Balci and GAO. TRAC-FLVN considers this aspect of models so critical that an organization specifically responsible for data/database management has been created in their combat model development directorate. They provide data to support the model developer's specific requirements. This frees the modeler from having to procure data for a model under development and causes data to be examined twice before use with a model. Input data management is an area with great potential for positively affecting the accurate assessment of future complex simulations' credibility.

Increased confidence in input data quality will increase the credibility of results achieved. TRAC-FLVN is also using operational graphics to help verify some models. In the training models directorate, the Joint Exercise Simulation System (JESS) is an example. An exercise designed to provide training to the command and staff structure of a maneuver battalion, it relies heavily on the use of computerized situation maps. As the development of JESS progresses, these same graphics are being used by modelers and programmers to identify anomalies in the execution of the simulation. The same type of graphics could be an outstanding verification tool for a modeler working with a higher-resolution combat model. CAA is familiar with and capable of applying all of these verification techniques to complex simulation models. Since this phase of assessing models is straightforward, CAA must ensure that a system of checks and reviews is in place so that verification is actually initiated early in the model's development.

E. Validation

The validation of complex combat simulation models is normally a very difficult process. Quantitative analysis of their results often cannot be conducted because detailed and reliable data from the real world system we are trying to model (i.e., modern combat) is generally not available. The usual compromise is to perform some subjective tests and evaluations in order to establish the validity of the model's output. The results achieved are commonly along the lines of being able to please some of the people some of the time but never all of the people all of the time. Some of the more common validation techniques are:

1. Face validity
2. Event validity
3. Model assumption validity
4. Comparison with historical data
5. Comparison to other models
6. Consistency checks
7. Extreme condition tests
8. Sensitivity analysis
9. Turing tests (Law and Kelton, 1980)

Face validity is the examination of the construction of a model for reasonableness by the model user and other knowledgable persons. Model ouput can also be examined for the appearance of reasonability. Sometimes called a peer or expert review, face validity techniques are the most frequently used validation technique among military agencies. An Air Force agency at Maxwell AFB in Alabama, the Air Force War Gaming Center (AFWGC), is unequivocal in its emphasis on expert review as the key to validation. Numerous general officers are brought in to officially review the results generated by simulations in production or development. The AFWGC also emphasizes as a credibility measure whether or not lower ranking officers believe the model's output is realistic. Both groups' comments and observations are used to calibrate the model until the results are considered valid.

The GAO report published in December 1987 cited peer reviews or "study advisory groups" as "the principal oversight and review body ensuring quality and consistency in the models when they are used in TRADOC's studies". Although the report praised the Army and TRADOC for

having established some formal procedure for model review during development, it described the overall effectiveness of these groups as "hampered....because it is not ultimately responsible for.....quality" and "encumbered by the large number of members" (DoD Simulations, 1987). Further, these groups have little decision making authority with regard to the model and are a logistical burden for the agency conducting the study. The GAO (and a TRADOC analyst workshop conducted on the subject) recommend forming small working groups of senior analysts who would meet periodically throughout a model development project to conduct critical reviews in depth, study problems as identified by the study groups and make decisions towards implementing solutions. This authority to make decisions based on the work of the study groups was considered the critical aspect of this idea.

The reviews of models conducted at CAA already use senior analysts as the key participants. A series of these reviews were analyzed for structure and methodology as part of this research. The results are summarized in Appendix B to this report. In general, the reviews were very informal and involved just a few senior analysts. No specific format was consistently used to prepare the reviews although the examination and discussion of algorithm structures was the most frequent approach taken. As an example, in a series of CAA documents reviewing AFP methodology dated from November 1986 to March 1987, several discrepancies in algorithm results are identified, expert opinions documented, and comparisons made between the CIP's generated by several different models. As a result, four specific alternatives for resolving the AFK vs. AFP controversy are presented in a memorandum to the director. Clearly, this is the type of process the GAO envisions.

However, some changes in the CAA procedure could strengthen it. These reviews need to be structured. A well documented peer review performed by lower level analysts, some involved and some not involved in the study, prior to the senior analyst review could identify problem areas and minimize the time needed to make necessary decisions. The decision authority level needs to be at the senior analyst conducting the review. The lower level peer review would be the appropriate group to look at the results of sensitivity analysis and extreme condition runs performed by the modeler for validation.

Comparing a model under development to an older, accepted model is a powerful approach which is not utilized often. The reason this approach can be powerful is that the old model acts as the real system being modeled and can provide comprehensive data for a thorough, quantitative analysis. Thus, statistical tests can replace some subjective tests. Even if the old model is not considered completely acceptable, the output of those portions or modules in which confidence is high (i.e., the direct fire module) can be used to validate the corresponding modules of the new model. Then subjective tests can be used on the remainder of the model. The training model directorate of TRAC-FLVN recently used this approach to validate the indirect fire module of JESS. The indirect fire module of the analytical model Vector-in-Commander (VIC) was used as the benchmark for the analysis. Using graphical and parametric techniques, the fractional damages achieved by JESS were calibrated to within 10% of VIC for a variety of scenarios (Flanagan, Asbury, Lewis, Venne, 1988). Comparing model results to historical data is usually accomplished by replicating a well documented battle from the past. The advantages of this technique are similar to comparisons with another model. CAA is

currently using VIC in a comparison with the recorded results of the Battle of the Bulge. The value of such a comparison is only as good as the quality and depth of this recorded data. Confidence in the data must be high. Several of the larger tank battles of the Arab-Israeli conflicts have been used to perform such comparisons. For instance, a 1983 study project group at the Army War College examined the 15-18 October period of the 1973 Arab-Israeli War when Israel moved to cross the Suez Canal into Egypt. Historical data was compared to that generated by an interactive, theater-level gaming model used at the College. Orders were issued to units exactly as it is known they were issued during the actual battle. Although this group was not experienced analysts, they encountered and documented problem areas commonly seen by analysts working on similar validation efforts. Historical events cannot be exactly replicated by a computer simulation. How closely must the simulation's results match the historical data in order to be useful? Unless the answer to this question can be quantitatively described during model development, a comparison with a historical conflict will be of little value (Baisden, et al., 1983).

An innovative technique which exploits many of the strengths of direct comparisons was discussed recently in a paper published by TRAC-MTRY entitled "Realism in Combat Models? Model Validation Using Results From Near Realistic Combat Scenarios". It advocates the use of data that is being electronically collected from near-real combat scenarios conducted during instrumented brigade-level exercises at the National Training Center (NTC) in Ft. Irwin, California (Galing, Wimberly, 1988). The authors describe the data collected at the NTC and present a convincing argument that despite certain limitations it can be

a very valuable tool in combat model validation. The training scenarios conducted at Ft. Irwin are as near combat as any in the world and the instrumentation of the exercises provides detailed and accurate data. Data from several brigade rotations through NTC was compiled and used to validate the JANUS combat developments model currently being used by TRAC-White Sands. JANUS is a graphics intensive, interactive model which allows the user to issue tactical orders and objectives through workstations. It was run with exactly the same objectives, orders and decisions issued during battles at the NTC and over the same terrain. Significant differences were found in unit movement rates and in the numbers of target engagements when JANUS and NTC data was compared. In general, JANUS battles led to higher movement rates and more engagements. These observations led to the discoveries that the JANUS algorithm for movement was incorrectly modifying unit rates of advance over hilly terrain and that JANUS did not account for "mal-engagements" (targets engaged out of range) which occurred significantly often during NTC battles. The paper concluded that such near-real combat data can be useful in validating combat models. With the establishment of a training center similar to the NTC for Europe and Korea based forces, the availability of such data for a variety of tactical situations could be plentiful. Modelers must pursue this resource to ensure their needs for credible data are understood and met. This data could be used to partially overcome the traditional lack of reliable war data for statistical model validation.

Another project is underway which continues in the direction of this research. It involves participants from TRAC-MTRY, TRAC-WSMR, AMMO, LLNL and the Naval Postgraduate School and is under the direction of

Professor Lester Ingber of the Physics Department of the NPS. The objective of the project is "to validate JANUS against NTC data" and to evaluate "the consistency and utility of NTC data" (Ingber, 1988). This project has the potential to uncover more validation-related uses for the NTC data and to accelerate its availability to analysis agencies. A final report is due in the Fall of 1988 and should be examined for applicability to CAA model testing.

Data from the NTC has also been used to conduct Turing tests with combat development models. TRAC-MTRY has surveyed combat leaders of units which have rotated through NTC training and of units stationed at the NTC which play the opposing forces (Soviet) during exercises. The survey gave a brief but concise tactical summary and battle losses experienced by both forces in a brigade or task force level battle generated by either an actual NTC battle or by the JANUS simulation. In summary, those surveyed were unable to distinguish between data generated by the NTC or by JANUS. The percentage of individuals correctly identifying actual NTC battle information ranged from 20 to 36% over five sets of battle data. A nearly identical range of 19 to 36% correctly identified JANUS generated data with over 50% answering "don't know" to the question of source of the data. In fact, in two of the five data sets more respondents selected the JANUS replication as the actual NTC battle. This fact alone is valuable information to an analyst trying to assess a model's credibility.

CAA's primary validation tool is currently the model reviews conducted by senior analysts. These reviews have generated valuable insight and should be continued but their structure should be formalized. The structure can be modified as experience is gained. Lower level

analysts should be involved. Direct model (or module) comparisons can be performed as applicable. The development of the use of NTC data (or other near-real combat data) for supplementing combat model validation should be monitored or even pursued directly by CAA.

F. Other Credibility Factors

As previously discussed, there are other factors which affect model validity in addition to the traditional verification and validation concerns. Many agencies have initiated programs for testing and controlling the quality of data input into simulation models. In fact, TRAC-FLVN has reorganized their combat model directorate in order to form a section specifically responsible for input data.

Several methodologies (i.e., that of GAO and that of TRAC-FLVN) for measuring model credibility included a planning and review step prior to actual coding of the model during which theoretical assumptions associated with the system being modeled are compared with those made in constructing the model. The usual tendency to begin coding as soon as the problem has been defined must be resisted until this planning and problem formulation is accomplished and documented. Although this is certainly a subjective measure, the model developer along with the coder and some sort of reviewing senior analyst should be able to come to an agreement that this has been accomplished. At CAA, a representative from the Math/Stat team should be included in this decision. The feasibility of solving the problem through simulation should also be addressed. The key is not to allow model construction to begin until the

senior individual is satisfied that sufficient time has been spent in formulating the real life problem.

A close examination of the exact purpose or definition of a model should be performed. All credibility measures should focus on this exact purpose. The level of effort required to establish model credibility should be directly related to the stated model purpose. Determining this level of effort must be done early. As pointed out by Landry's model confidence specification matrix methodology, the validation of a model has no meaning outside the "context of the model's purpose" (Landry, 1983).

Several Army agencies are now assigning configuration control responsibility for models under development. This includes standardization of the model for export to other users and controlling or approving authorized changes or versions of a model in order to eliminate the proliferation of undocumented or altered versions of the model. The GAO report praised the Army's Model Improvement Program (AMIP) for assigning "responsibility for the control of the model's configurations." (DoD, 1987).

Finally, the strengths and weaknesses displayed by models or sub-modules of a model must be fully documented. As discussed in the analogies section of this paper, a summary of strengths and weaknesses should be included in any military model catalog. Modelers could then more intelligently use such catalogs when attempting to use already developed models in later projects. Many of the subjective opinions of analysts involved in the model's development could be recorded in the actual model documentation. This will allow these models to be used intelligently later to develop or even test new models or modules. The

GAO report stresses full disclosure of model testing and performance in order to avoid future duplication of errors or marginal performance.

G. Applications to Complex Military Simulation Models

Some of the methods available for model credibility assessment are directly applicable to complex military models such as those developed at CAA. Others must be modified if they are going to be used effectively with large models while some methods are simply not able to be used in this situation. CAA must develop a methodology for model credibility assessment which includes methods which fall into the first two categories.

Documentation of code and structured programming practices are model verification methods which are directly applicable to complex military models and may be even more critical in this application due to the large amount of code involved. CAA must ensure programmers are following these practices. However, methods normally used to check the quality achieved by programmers may not be practical with large models. For example, a structured walk through using a group of programmers, model developers and model users to review code and documentation line-by-line would be very time consuming. As described in Section VI of this paper, a random sampling scheme of checking small sections of code throughout the program would be more efficient and theoretically just as effective in measuring the quality of documentation, structure and correct functioning of code. The sampling scheme must be constructed in accordance with standard acceptance sampling techniques. The Math/Stat team should be responsible for the sampling design and definition of acceptance/rejection criteria.

Clearly, a trace of all model events is not a feasible verification method for a large model. However, a trace of a few key events over time could be done. A trace of many events over a very short time period or short section of code is also feasible. For example one unit could be traced through time. Other approaches to conducting a "partial" trace include tracing all units that pass particular points in the model or tracing all model activity between two occurences of a particular model event. By allowing the analyst to define and select these modified traces, this method can be applied to large models.

The planning and review phase used by TRAC-FLVN is directly applicable to complex models although it may need to be conducted at the level of program modules due to overall program size. Examination of the exact purpose of the model should be accomplished during the planning and review.

Input data review must be automated to be applied effectively to the large databases needed by CAA. Until this automation can be developed, this method of verification is not directly applicable. Perhaps this automation can be developed sequentially for different categories of input data so that quality checks of data can be gradually implemented. The quality control technique of control charting discussed in Section VI could be used in this implementation. For example, develop and implement control charts for unit movement related variables and once functioning repeat the process for another group of variables.

Operational graphics used as a verification tool have to be employed over a smaller, representative portion of a theater-level model. For example, graphically portray unit movement, attrition and disposition

over a division sector assuming it to be representative of the model's behavior over the entire theater.

Face validity is directly applicable to complex military models. Reviews of models in development should continue at CAA. The exact format of these reviews must be reconsidered as previously discussed in the validation portion of this section.

Event validation is comparing model generated events to real world events. It is too extensive for a large model and actual events from the real system (combat) may be theoretical in nature. As with event traces, selected events could be examined for validation.

Validation of a complex military model through comparison to an older model is difficult due to the size of the models involved. However, comparisons would be appropriate to validate modules of models being constructed. If a particular module of an existing model enjoys high credibility, statistical comparisons of selected outputs can be used to validate or calibrate the newer model's module. TRAC-FLVN's validation and calibration of the indirect fire module of the JESS model is an example.

Comparisons with historical battles are applicable if differences in modern capabilities are considered. Quantifying these differences could be very time consuming but if several battles were selected as 'benchmarks' and used repetitively for model validation the effort could be worthwhile. One such historical comparison should be done per complex simulation developed. Less than satisfactory results for important variables would indicate the areas of the model which need further testing and review.

The use of near-combat data, such as NTC data, for statistical comparisons and validation is not applicable primarily because NTC or other realistic data is not currently available in a form usable by analysts. Data being generated and compiled at the NTC must be reviewed and tested by modelers to determine how it can be best be used. Data from NTC used by TRAC-MTRY had to be reformatted before it was used to validate portions of the JESS model. Data collection at the NTC can be modified to provide more usable data to analysis agencies. The validation of the movement and engagement algorithms of JESS is an example of how such data could be used as a validation tool. Results of on-going studies of NTC data, such as that headed by Ingber, must be considered and implemented as they become available. Any use of near-combat data for direct comparison to model output will increase confidence in an evaluation of model credibility.

Consistency checks are described as the examination over time of a model to ensure it continues to adequately describe the real world system (Banks, et al., 1986). Changes in doctrine and equipment capability are examples of areas which require periodic updates. This cannot be done on a continuous basis for a large complex model. A logical modification to this method would be to require an annual review by an analyst and written certification of model currency.

Both sensitivity analysis and extreme condition testing of complex military models are hampered by the number of variables involved. A modification to both methods so that they could be used in validation is to group important variables into a small number of topical sub-groups. For example, group all variables related to manpower (or firepower, logistics, mobility, etc) together. Vary the values of these variables

as a group. Small changes in these variables which favor the blue force would be expected to tip model results towards them in a reasonable way. Large changes should cause the model to "blow-up" in some logical fashion. Grouping variables will reduce needed model runs to a reasonable number. Each group of variables should be tested separately.

Establishing configuration control responsibilities in CAA and full documentation of model strengths and weaknesses, both recommended by the GAO, are directly applicable to complex military models and should be accomplished. Appropriate sections/directorates should be assigned configuration control for models developed by CAA. They would maintain the "official" version of the model. Final model documentation should require a complete description of strengths and weaknesses.

A summary of applicable techniques available to CAA for model verification and validation indicates a plentiful supply for the development of an effective methodology as shown in Exhibit VI-2. Future progress in the areas of automated input data checks and comparisons with near-combat data could add strong statistical tests to any methodology based on the above techniques.

H. Conclusion

Verification and validation of a complex simulation model are only two aspects of assessing the model's overall credibility level. Other factors such as input data quality, model purpose and model characteristics critically affect the performance of the model. Any agency testing a model in development should use as many of the existing verification and validation methods as possible in evaluating its performance. Many of these techniques are being used in innovative or

Exhibit VI-2. Summary of Applicable Techniques

Method	Applicable	Applicable w/Changes	NA
Documentation	X		
Structured Programming	X		
Verification of Documentation		X	
Verfication of Code		X	
Traces		X	
Planning/Review Phase	X		
Input Data Review			X
Operational Graphics		X	
Face Validation	X		
Peer/Expert Review		X	
Event Validation			X
Comparison w/Model		X	
Comparison w/Historical Data		X	
Comparison w/Near-combat Data			X
Consistency Checks		X	
Sensitivity Analysis		X	
Extreme Condition Tests		X	
Configuration Control	X		
Documentation of Model Strategies/Weakness	X		

unusual ways by military analysis agencies. These ideas can be implemented in testing of models currently in development.

CAA's stated goal is to implement a more formal verification and validation program. Techniques must also be developed to assess any other factors considered important to the model's credibility. Techniques which are directly applicable to or can be adapted so that they are applicable to large, complex military simulations are available. Some or all of these must be adopted and/or revised in order to give CAA a methodology for determining whether or not a model is ready to move from the development phase to the production phase. Testing must begin early in the model's life cycle and responsibility must be assigned for ensuring that it continues throughout the model's life. The Math/Stat team would be a logical choice at CAA to become involved in such a quality control effort.

REFERENCES

Baisden, E., et al., (1983), "Validation of the USAWC Student War Gaming Model," US Army War College, Carlisle Barracks.

Balci, O., (1986), "Credibility Assessment of Simulation Results: The State of the Art," Technical Report, Virginia Polytechnic Institute, pp. 2-24.

Banks, J. and Carson, J., (1984), Discrete-Event System Simulation, p. 379, Prentice-Hall, Inc., Englewood Cliffs, N.J.

Banks, J., Gerstein, D. and Searles, S., (1986), "The Verification and Validation of Simulation Models: A Literature Review," School of Industrial Engineering, Georgia Institute of Technology, Atlanta.

"DOD SIMULATIONS: Improved Assessment Procedures Would Increase the Credibility of Results," (1987), GAO/PEMD-88-3, Washington, DC.

Flanagan, S., (1988), "Verification and Validation: A TRAC Approach," Phalanx, Washington DC.

Flanagan, S., Asbury, T., Lewis J., Venne, T., (1988), "JESS 1.0 Indirect Fire Validation," Technical Memorandum TRAC-F-TM-0288, Ft. Leavenworth.

Galing, B., and Wimberly, B., (1988), "Realism in Combat? Using Results From Near Combat Scenarios," p. 4, Technical Report, TRAC-Monterey.

Ingber, L., "Mathematical-Model Validation of Combat-Computer Models: Comparison of Janus Against National Training Center Data," Memorandum dated April 18, 1988, Naval Postgraduate School, Monterey, California.

Landry, M., Malouin, J., Oral, M., (1983), "Model Validation in Operations Research," p. 207-220, European Journal of Operational Research, 14, Holland.

Law, Averill M. and Kelton, W. David, (1982), Simulation Modeling and Analysis, p. 333-334, McGraw-Hill Book Company, New York.

Schlesinger, S., (1979), "Terminology for Model Credibility," p. 103-104, Simulation No. 32.

VIII. Recommendations

A. General comments

It is recommended that CAA structure verification/validation efforts around the conclusions and observations set forth in Sections IV through VII. The basic CAA methodology should be structured around the seventeen step process outlined in part H of Section V (conclusion to large systems). During the initial planning of a verification/validation effort on a particular model (steps 1 through 10), the inferences from the analogies (part D of Section IV) should be considered as general "good practice" guidelines in structuring the verification/validation effort. The statistical methods outlined in Section VI should be employed whenever possible when checking the model for operational validity (step 16). Section VII outlines potentially useful techniques for steps 11 through 15.

B. Specific comments

1. Section IV: Analogies.

Each analogy offered several constructs that could be employed in the verification/validation of complex simulation models. The most important ones are considered here.

a. A model's history of validity and utility is an indicator of future validity and worth. A model's future worth to CAA should be evaluated prior to launching an expensive verification/validation effort.

b. Normal ranges (benchmarks) should be established for model parameters and variables. The model code should be modified to call attention to parameters and variables that fall outside of accepted benchmarks during simulation.

c. If the verification/validation effort is hampered by time or funding shortages, the most critical modules and subroutines should be thoroughly examined with specific tests and standards in mind.

d. Peer reviews should be conducted by analysts who do not have a vested interest in the model. The peer review process should be thoroughly structured and specified by CAA.

e. The verification/validation process should be thoroughly documented throughout the effort.

f. A central authority must monitor and control the verification/validation effort. All changes to a model should be implemented only after thorough testing and with the approval of the central authority.

g. Models must be evaluated with respect to specific performance objectives. These objectives must be specified at the start of the verification/validation effort.

2. Section V: Large Systems.

Section V establishes a seventeen step process for verification/validation of CAA models. First, CAA must establish a standard operating procedure that defines the conduct of the verification/validation process

within the agency. The exact performance requirements and output requirements for the model must be examined, redefined if necessary, and thoroughly specified prior to initiating the verification/validation process. It must be understood that the credibility of a model cannot be established with 100% accuracy. Instead, an "accuracy criterion" should be established for each model. This criterion should specify the level of accuracy required by the model to instill confidence in the decision maker who uses its output. The level of verification/validation should be established for each module or subroutine in the model. Extensive effort should be allocated to the most critical portions of the model. The model's performance and operational requirements should be thoroughly specified in a "requirements" document that is constantly referenced during the verification/validation process. This document serves as a road map to ensure that all critical aspects of the model's review are accomplished. Once the specific verification/validation requirements of a model are specified, the specific techniques of reviewing simulation development (step 11), conceptual model assessment (step 12), software verification (step 13), operational validity (step 14), and data validity (step 15) should be reviewed for applicability. After a consideration of the resources available to apply them, the most promising techniques should be utilized. It is particularly important to constantly monitor the experience of actual model users. This feedback provides constant information on the daily performance of the model and identifies potential problems with the model as they occur.

3. Section VI: Statistical methods.

Statistical methods can be useful in increasing the understanding of a model and obtaining some level of confidence that the model is correctly imitating the simulated system.

a. Control charts are applicable to tracking the parameter values of a simulation model. They require few assumptions and their use is restricted only by the analyst's imagination. They do not require replications and do monitor the statistical mean of important parameters, which should increase understanding of the model's characteristics and trends. They should be applied to the analysis of input data, model processes, and model output.

b. Acceptance sampling provides useful techniques for drawing inferences about the accuracy of a module based upon the examination of only a portion of module code. This technique maximizes the limited time and resources available at CAA.

c. Fractional factorial experimental designs allow the CAA analyst to determine the significance of several factors (parameters or variables) in an algorithm or module while performing a minimum number of replications. Hence, they also maximize the available resources of a limited budget.

d. Cluster analysis allows the CAA analyst to reduce a large number of data elements into a form that highlights the relationships between the data elements. Hence, understanding of data relationships are enhanced. Cluster analysis algorithms can be applied to several model outputs generated from simulation runs with different inputs. This

procedure assists the analyst in determining model response to various input modifications.

4. Section VII: Extension of known verification and validation techniques.

This section describes the concept of model credibility to include the traditional definitions of verification and validation. It identifies a number of traditional techniques that can be extended to CAA models.

a. Comparing portions of a new model to the corresponding modules (those in which confidence is high) of an older model can provide useful results. The old module serves as the historical data base upon which the performance of the new module is compared.

b. The use of automated input data review would considerably lessen the amount of work required by a CAA analyst to validate large model data bases. Continued development of this capability at CAA should be considered.

c. Operational graphics simplify the understanding of internal model action by providing "snapshots" of the model at discrete points in time. CAA should incorporate the use of these graphics in current models. CAA should pursue the synthesizing of near-combat data from the NTC into a useable form for use as a comparison output data base for model outputs.

e. CAA should establish configuration control guidelines and documentation guidelines for all models in use and under development.

Comparison to Other Models - Various results of the simulation model being validated are compared to the results of other valid models. By using similar previously validated models it may be possible to establish a level of credibility for particular functions of the model in question.

Consistency Checks - This refers to examination of the simulation model over time to insure that it continues to adequately describe the real-world system.

Documentation - This is the recorded information concerning a model. Documentation can be sub-divided into two parts; descriptive and technical. Descriptive documentation is general information about the model's theory, capabilities, limitations, and assumptions. Technical documentation is that detailed information which describes how the simulation model works and the exact mechanics of the model. This information is usually in written form.

Event Validity - The "events" or occurrences of the simulation model are compared to the real system to determine if they are the same.

Extreme Condition Test - The model structure and output should be plausible for any extreme and unlikely combination of levels of factors in the system. Also, the model should bound and restrict the behavior outside of normal operating ranges.

Face Validity - This technique refers to asking people knowledgeable about the system whether the model and/or its behavior is reasonable. Face validity can be used in all facets of the system model.

Historical Methods - The three historical methods of validation are rationalism, empiricism, and positive economics. Rationalism assume that everyone knows whether the underlying assumption of a model are true. Then logical deductions are used from these assumptions to develop the correct (valid) model. Empiricism requires every assumption and outcome to be experimentally validated. Positive economics requires only that the model be able to predict the future and is not concerned with its assumptions and structure.

Input-Output Transformations - This refers to the model's ability to predict the future behavior of the real system when the model input data match the real inputs and when a policy implemented in the model is implemented at some point in the system. The structure of the model should be accurate enough for the model to make good predictions for the range of data sets which are of interest.

Operational Graphics - The model's operational behavior is displayed graphically as the model moves through time. This helps to visualize the progress of the simulation during execution and aids in the detection of verification errors.

Sensitivity Analysis - This validation technique consists of changing values of the input and internal parameters of a model to determine the effect upon its output. The same relationships should occur in the model as in the real system.

Statistical Tests - This refers to statistical procedures used to determine input data validity as well as for comparing real-world observations and simulation output data.

Structured Programming - This refers to the use of specific techniques which enable one to more easily understand programs and troubleshoot existing code. Techniques include program modularity and top-down design. Program modularity is defined as the decomposition of the model into sub-modules which have well defined functions and interfaces. Top-down design refers to the notion of creating a detailed plan of the model before writing the code in order to avoid revising the original structure of the model.

Structured Walkthrough - This refers to assembling a group of model developers and users to participate in a line-by-line evaluation of the model or a sub-module of the model. A structured walkthrough is utilized to detect errors early in the model development cycle, to provide an informal review, and to encourage technical exchange in a constructive, non-fault finding atmosphere.

Traces - The behavior of different types of specific entities are followed through the model's execution to determine if the coding is correct and if the necessary accuracy is obtained. The trace is accomplished by printing out the state of the simulated system just after each event occurs.

Turing Tests - People who are knowledgeable about the operations of a system are asked if they can discriminate between system and model outputs.

Appendix B

PEER REVIEWS

1. Subject: Review of MICAF

Date : 13 November 1986

Purpose of study: Inform the director of problems encountered with the model

Methods used:

a. Algorithm critique:

(1) Distribution of type duels

(2) Half-life factor

(3) Vulnerability reflected twice

(4) Allocation of ranges among duels remains unchanged

(5) All weapons in a particular posture are in the same situation

(6) There is no suppression effects for any fire

b. Experts opinions - Blue CAS sorties underestimated for a typical division

c. Input analyzed - Target values are based on group judgement

2. Subject: Review of AFP Methodology

Date : 11 December 1986

Purpose of study: Supplement to 13 November study

Methods used:

a. Algorithm critique:

(1) Half-life factor

(2) Vulnerability reflected twice

(3) All systems have the same opportunity to engage in combat

(4) Allocation of ranges among duels remains unchanged

(5) All weapons in a particular posture are in the same situation

(6) No calibration for support forces

b. Experts opinions -

(1) CIPs clearly show a bias toward indirect fire systems

(2) Modulated COPs for divisions were too high relative to unmodulated

c. Compare to other models - Recommend that CIPs and COPs be compared to other models.

NOTE: In the directors response, he recommends using COSAGE results to calculate half-lifes for input to AFP. Asks what appropriate CAS levels should be and source of information. Questions some of the conclusions that algorithms are incorrect.

3. Subject: Review of AFP Methodology

Date : 17 December 1986

Purpose of study: Expound on two previous memos

Methods used:

a. Algorithm critique:

(1) CIP MOE is inappropriate for this methodology

(2) CS/CSS factors lack proper calibration to combat potential

(3) CAS not being given its proper role (too few CAS, and targeting)

(4) Division resources not clearly delineated from EAD resources

(5) Type duels not depicted over their appropriate frequencies

(6) One-dimensional CS/CSS factors aren't good representations of their contribution

(7) Target values are arbitrary

NOTE : Recommends getting target values from other models, using Eigenvalue method

4. Subject: AFP validation review, 4th memo

Date: 25 Feb 1987

Purpose of study: Inform the director of problems encountered with the model.

Methods used:

a. Algorithm critique - Half-life factor--believes that artillery and helicopters are not portrayed as well as direct fire duels

b. Model comparison - AFK, MERCAF, Eigenvalue, and AFP lead to different CIPs and different force improvements. COSAGE was also compared for some weapons.

NOTE: Analysts' comments at the end indicate that no one is sure which model to believe.

5. Subject: AFK vs AFP

Date: 2 March 1987

Purpose of study: Present alternatives to these two systems

Methods used:

a. Algorithm critique - more explanation of same topics as previous memo

b. Comparison to other models - same analysis as previous memo

c. Alternatives presented - Four alternatives are presented, from using different models to attempting to fix problems in current system.

6. Subject: Peer Review of WARF methodology

Date: 16 Feb 1988

Purpose of study: Large differences in WARFs in P93 and P90. Wanted to determine why these changes occurred and which was correct.

Methods used:

a. Documentation - out of date or nonexistent, had to rely on notes and verbal explanations

b. Algorithm critique:

(1) Non-modeled systems not handled correctly--CAPP no documentation

(2) Model has been updated since P90 so P93 results could be expected to be different

(3) War reserve items unconstrained from other resources

(4) Point estimates of WARFs developed

(5) Process of aggregation in going from COSAGE to CEM and disaggregation from CEM and COSAGE to WARF

c. Experts opinions - Historical factors applied to losses

d. Recommend comparing CAPP model with COSAGE output. Also recommend better documentation (audit trail)

e. Input - Needs to receive more attention.

NOTE: Following this peer review is a two page critique of WARRAMP methodologies. The critique addresses the same arguments as shown above--recommends comparison with other models and mentions experiments with COSAGE and CEM at half strength.

7. Subject: Validation of CFAW

Date: 20 May 1987

Purpose of study: Assist a study team in determining validity of model

Methods used:

a. Sources of information:

(1) Meetings

(2) Model documentation

(3) Internal study memoranda

(4) Previous applications

(5) Earlier studies

(6) Sensitivity analysis--of model and sub-models

b. Purpose of model emphasized. Also emphasized that one cannot validate a model by looking only at the inputs and results.

c. Model has players--increased variability. Recommend that other be studied to see if they need to be stochastic.

Replacement by deterministic modules will allow replication and understanding of the model to be increased.

d. Insufficient time for quality input

e. Model has never been validated by experienced users

f. Ground combat model algorithm is bad (only one posture, hex boundary, unit formation, rates of fire - are all weak)

g. Air battle model - radar not represented correctly, force multiplier inaccurate, and engagements not broken off realistically

h. No model documentation-only drafts

8. Subject: Special Review of the ARQ model

Date: Unknown--our copy is on briefing charts

Purpose of study: Determine if development of ARQ should continue and what should be done to fix problems

Methods used:

a. Approach:

(1) Learn about WARRAMP and ARQ

(2) Compare WARRAMP and ARQ methodologies

a. On paper

b. Run ARQ and ATCAL attrition equations

(3) Verify ARQ

a. Check files

b. Run ARQ for hand-calculable cases

(4) Define potential alternatives to ARQ

(5) Evaluate ARQ and the alternatives

b. Problems

(1) No detailed flow chart or documentation of ARQ

(2) No documentation on past verification and developers no longer at CAA

(3) Only one month--couldn't run enough ARQ-ATCAL comparisons

c. Gives history of model

d. Gives purpose of model--replace WARRAMP at 10 times the speed

e. If ARQ can come within 15% of WARRAMP-good enough. Lists other criteria for success

f. Discusses algorithm in depth-also those of its competitors

g. Gives comparisons to WARRAMP--not good

h. Says artillery losses are low--experts' opinions

i. Lists alternatives.

Appendix C

A SET OF TEMPLATES FOR EVALUATING WARGAMES

(BENCHMARKS)

Prepared by

ROBERT McQUIE

U.S. Army Concepts Analysis Agency

October 1988

E X P L A N A T I O N

A set of templates are presented here that can be used in comparing wargame results with history. These templates are based on data about 260 historical battles between 1937 and 1982. This data has been assembled over a quarter of a century by Colonel Trevor N. Dupuy, USA, Ret and his military historian colleagues at Data Memory Systems, Inc. This historical data, like wargame data itself is not easy to interpret.

To facilitate interpretation and enable the data to be compared with wargame results, 31 ratios and rates were extracted from it describing key aspects of each battle. The data and results may be found in a report published by the US Army Concepts Analysis Agency. The 31 measures summarize the following aspects of the 260 battles:

relative advantage
target density
weapon density
casualties & losses
movement & duration
artillery fire

From the array of values for each of the 31 ratios or rates, three values were extracted:

a. The median value.

b. The range in which the middle 50% of the values are found.

c. The range in which 90% of the values are found.

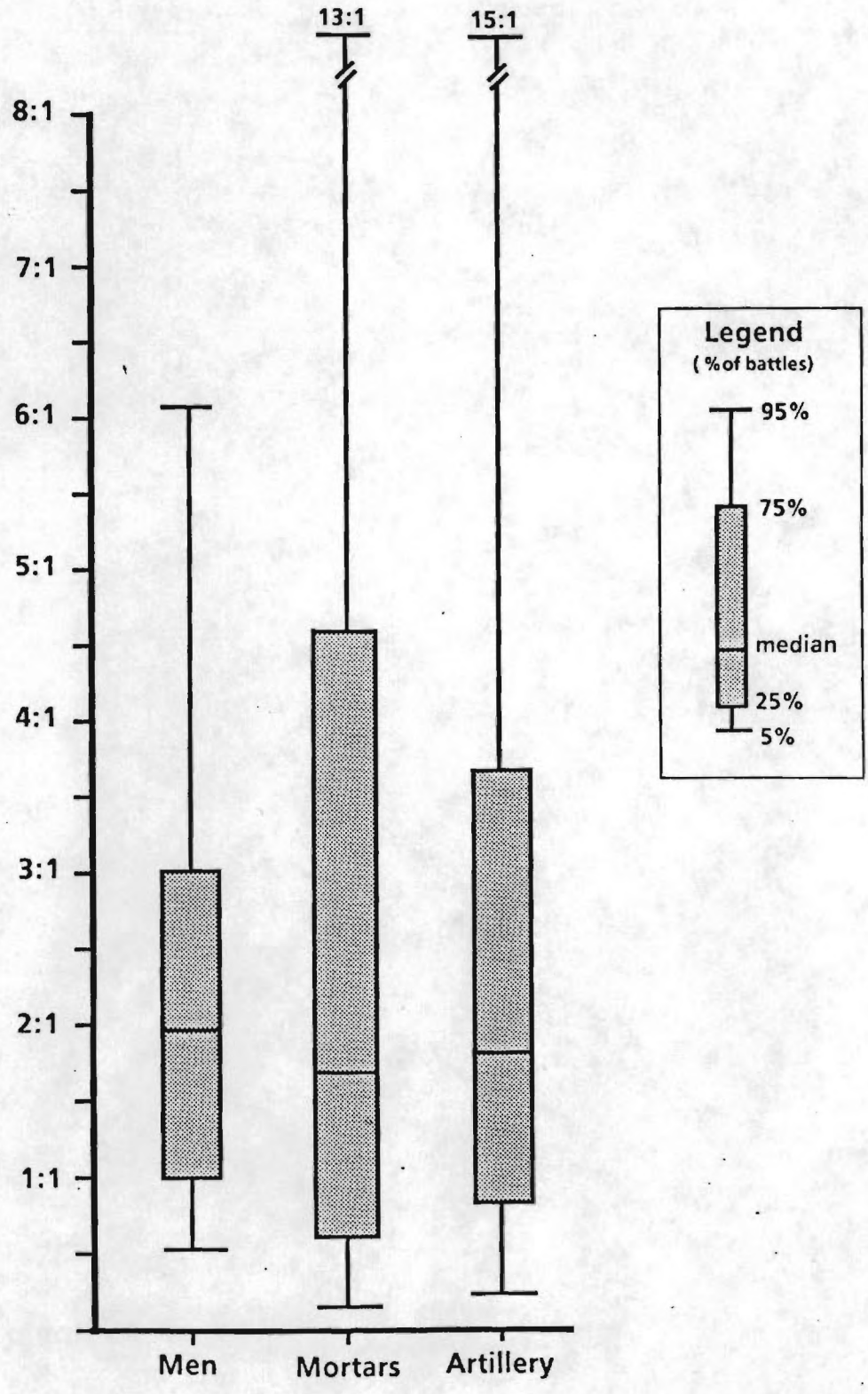
These three values were used to construct the templates on the following pages. In each template, the "box" shows the range in which the middle 50% of the values were found; the "whiskers" at each end of the box show the range of values extended to 90% of the battles. The median value of the measure is shown as a horizontal line across the middle of the box.

When the range of a characteristic is so wide that it cannot be diagrammed on the template, the whisker has been truncated, with the actual value shown by a small figure at its end. Otherwise, the scale at the left edge shows the values and units of measure. Questions about these templates may be referred to the author (AV: 295-5227).

The templates provide, it may be noted, a summary description of combined arms combat as it has existed in our lifetime.

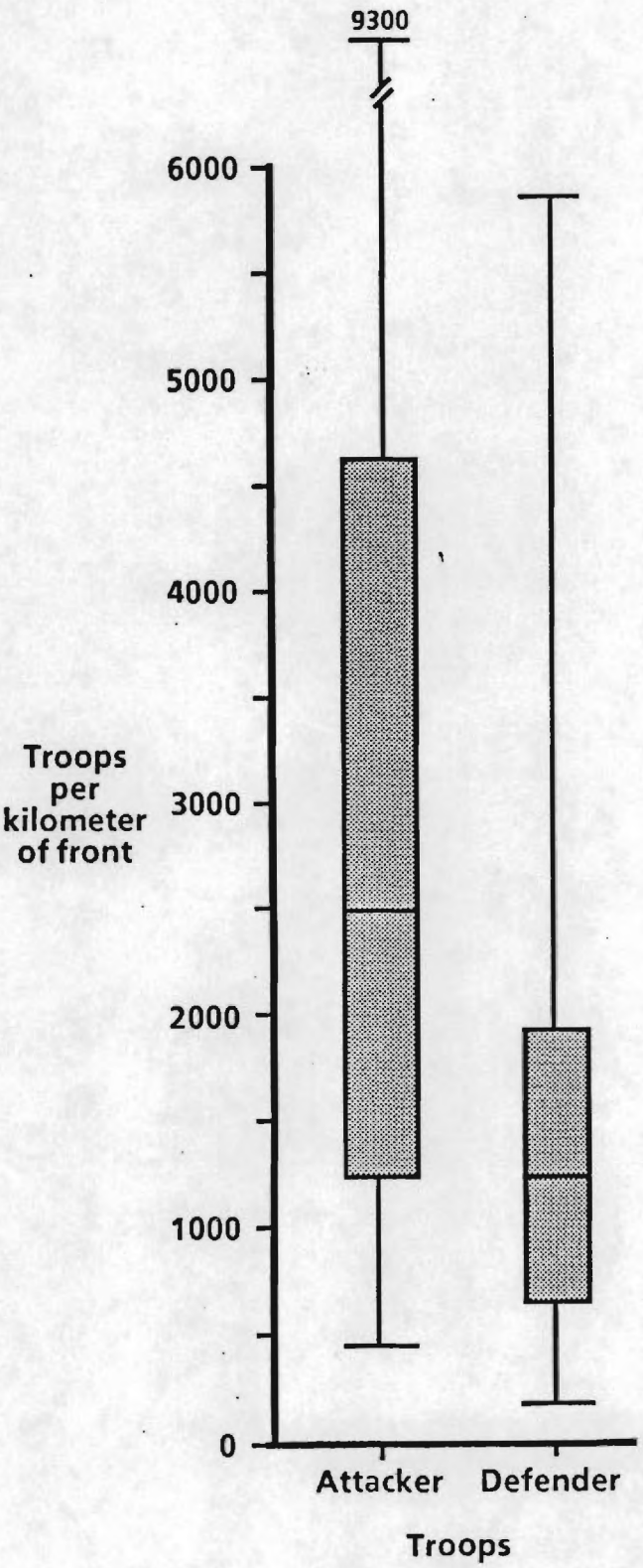
INITIAL CONDITIONS

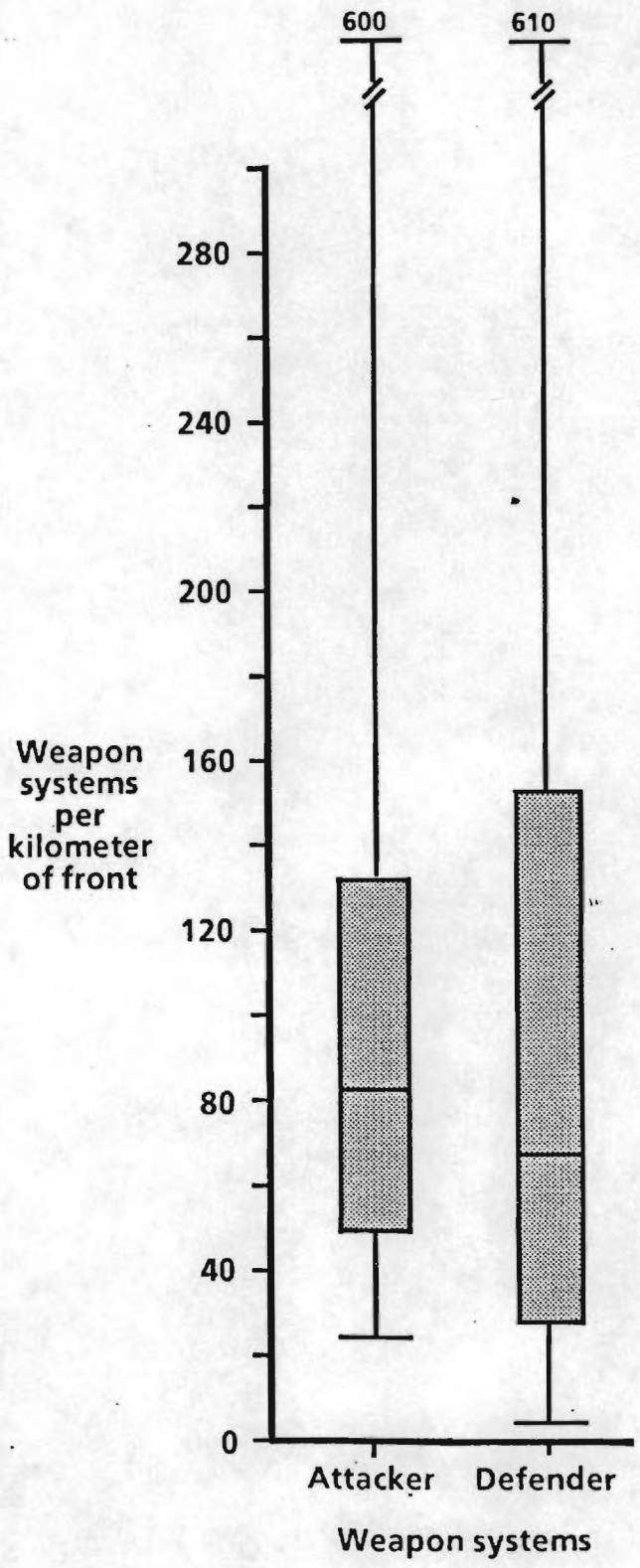
1. RELATIVE ADVANTAGE



INITIAL CONDITIONS

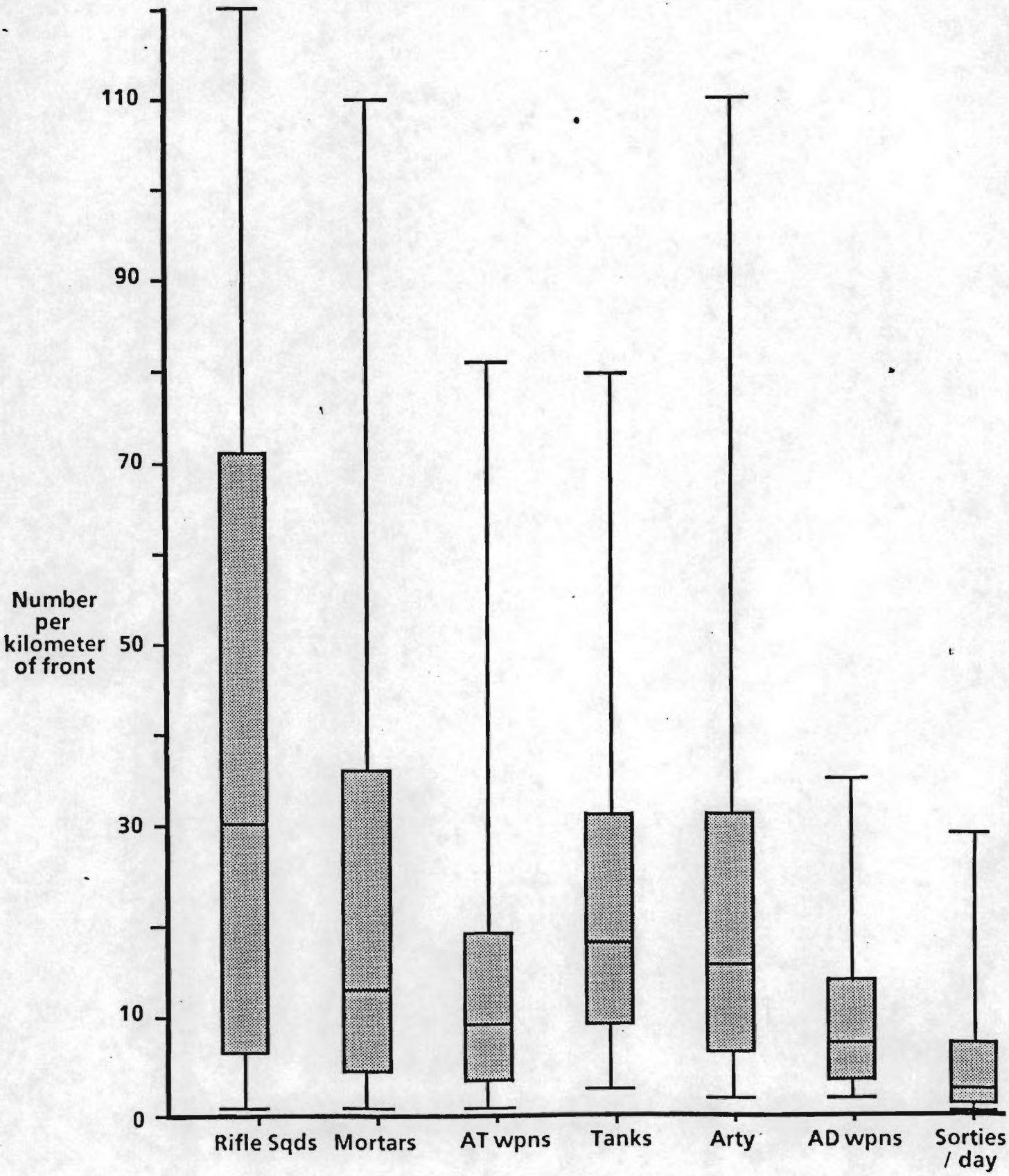
2. TARGETS





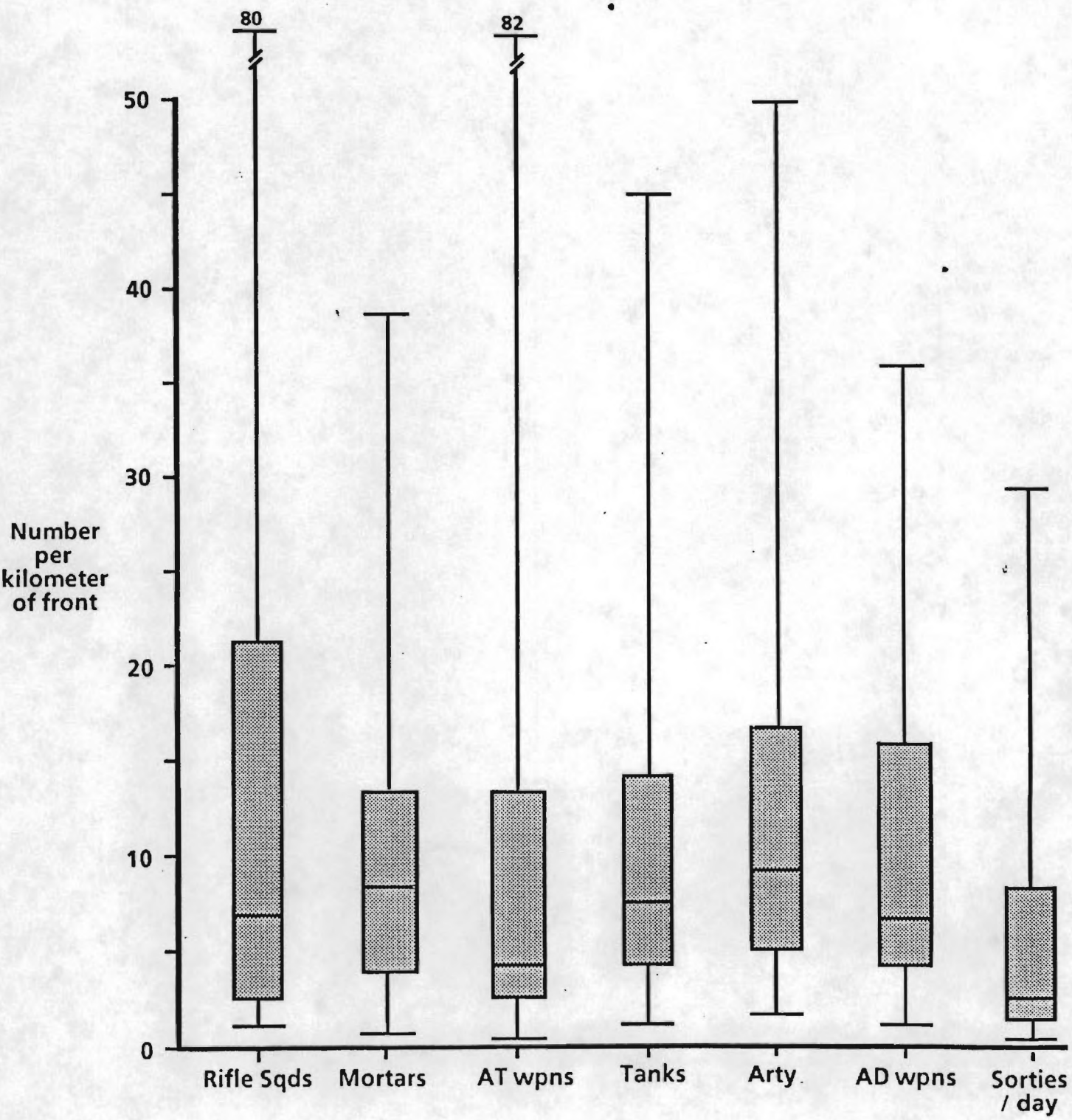
INITIAL CONDITIONS

3. ATTACKER WEAPONS



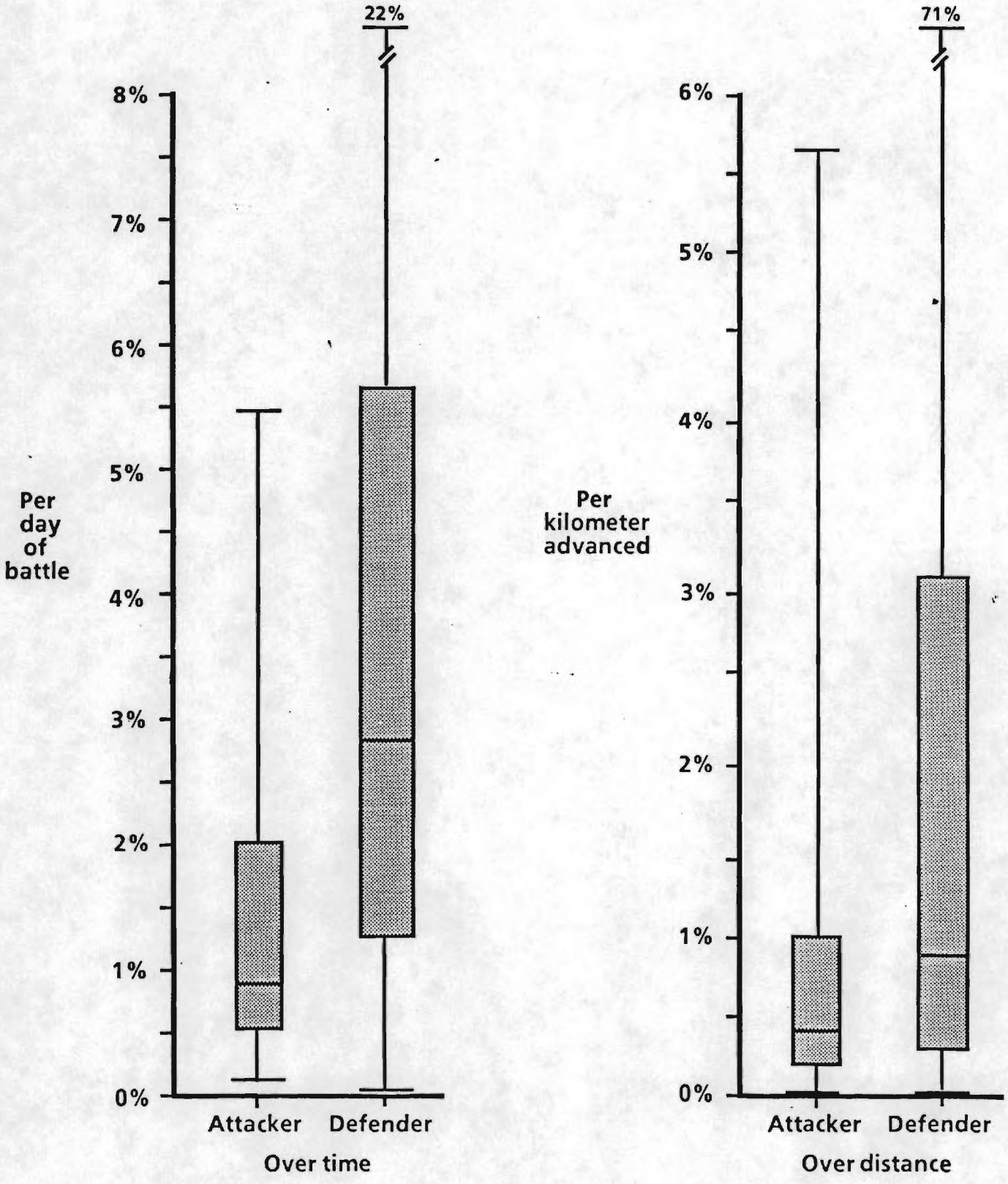
INITIAL CONDITIONS

4. DEFENDER WEAPONS



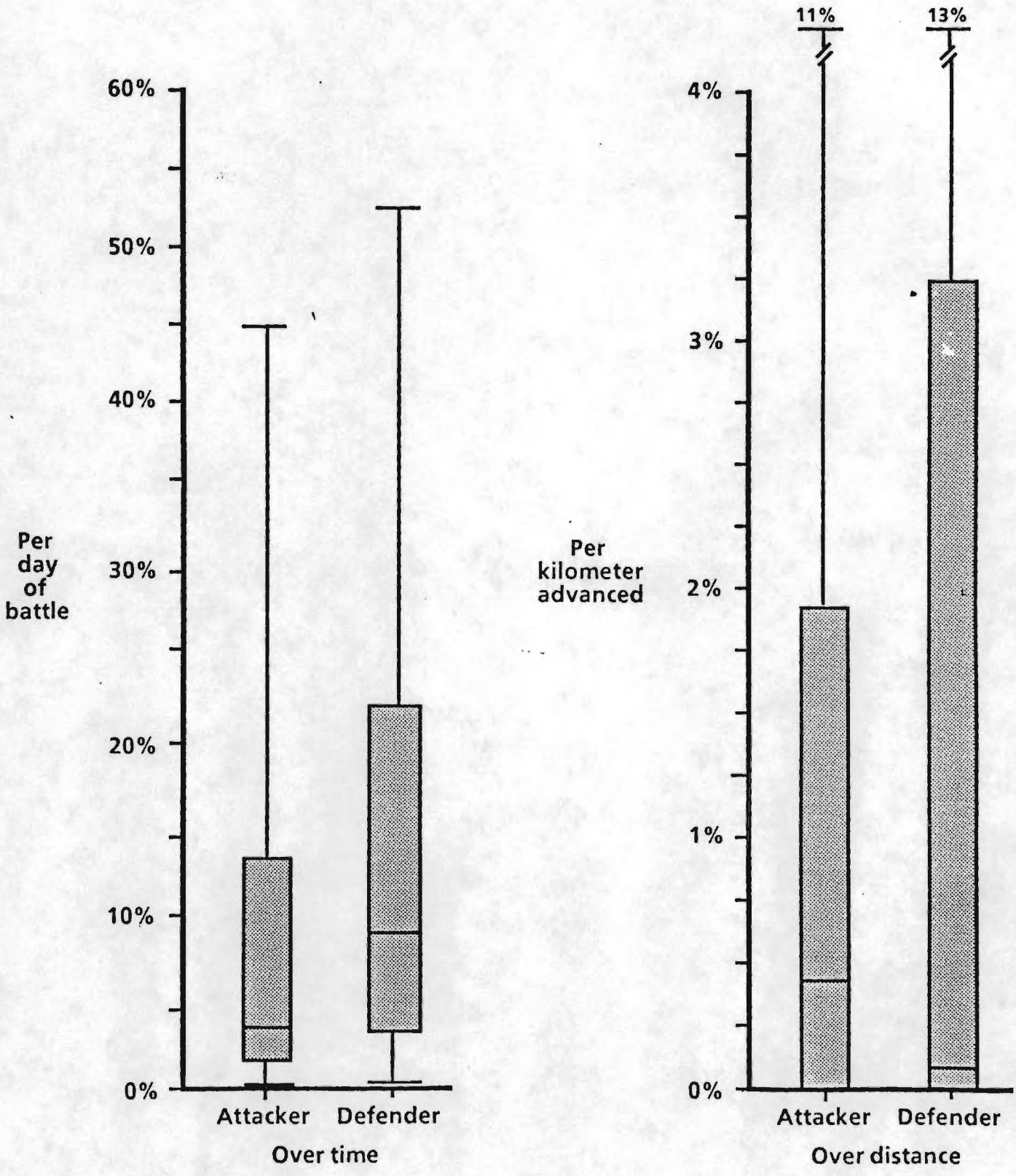
OUTCOMES

5. CASUALTIES



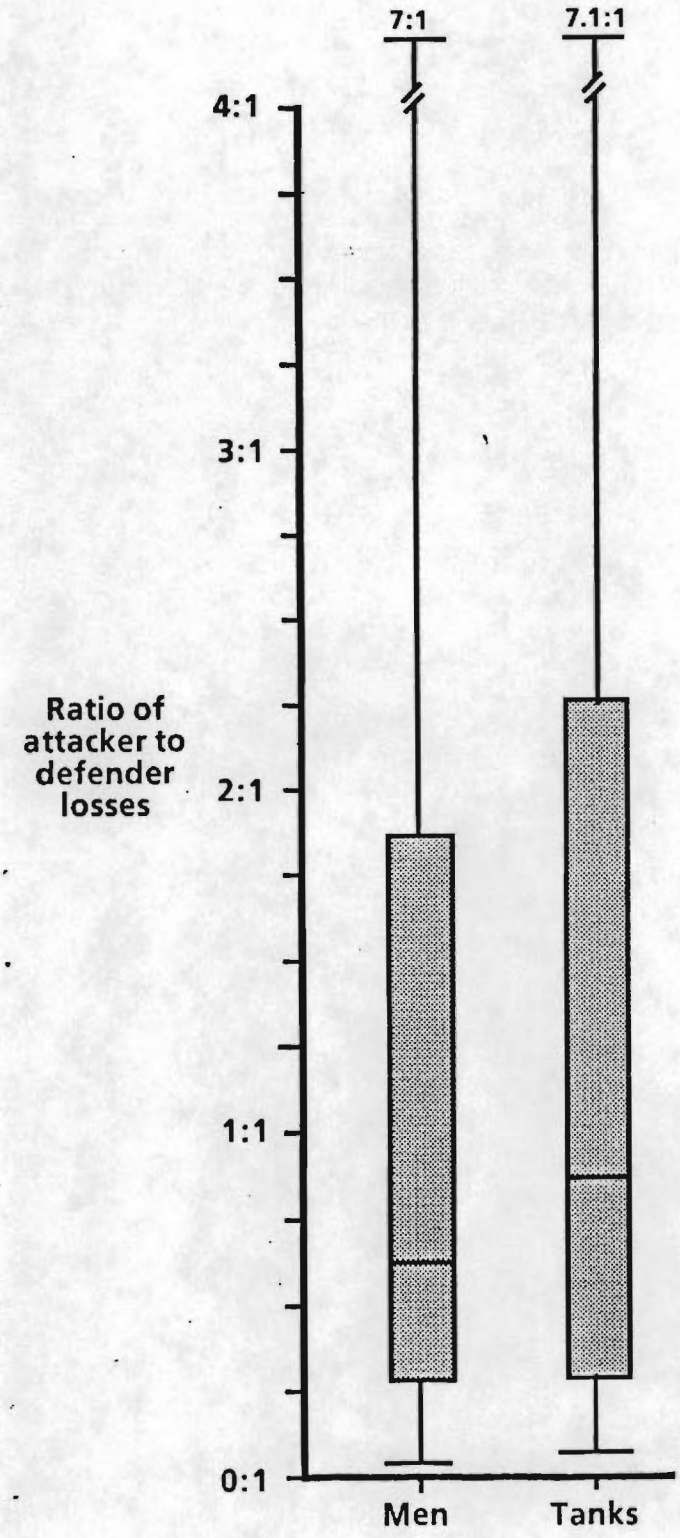
OUTCOMES

6. TANK LOSSES

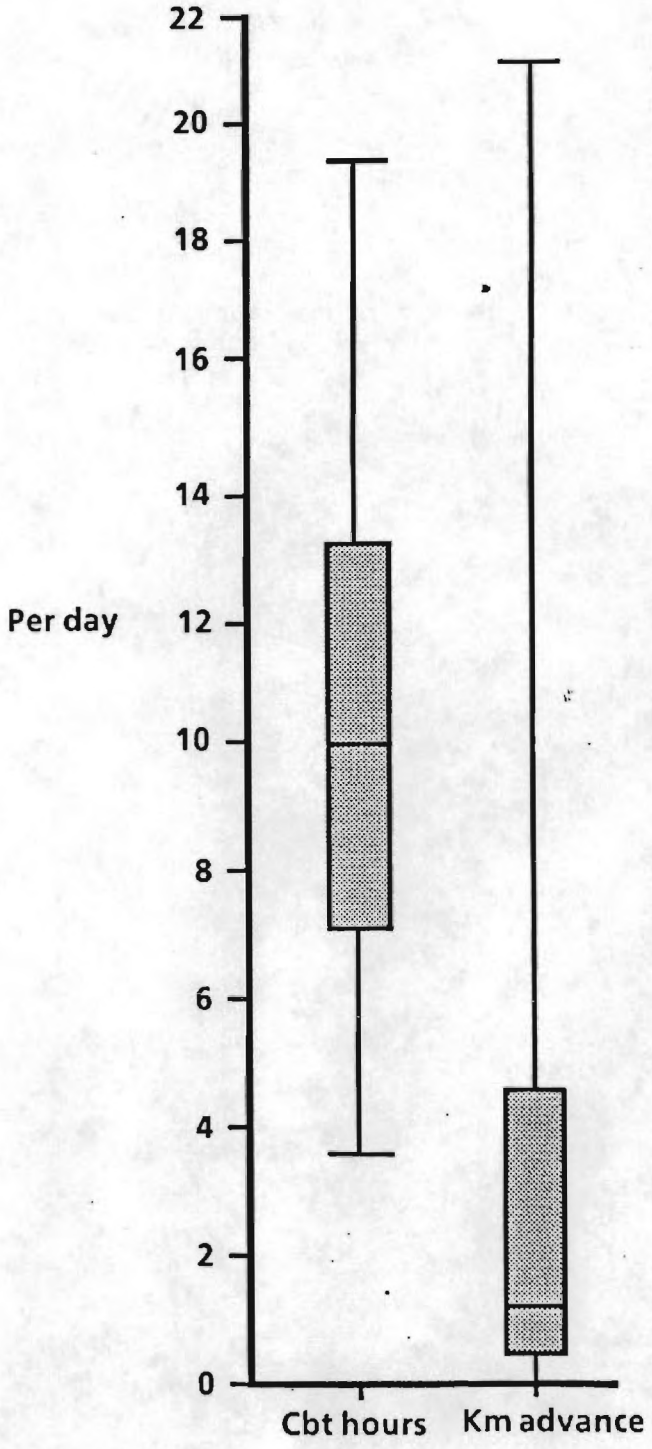


OUTCOMES

7a. LOSS RATIO

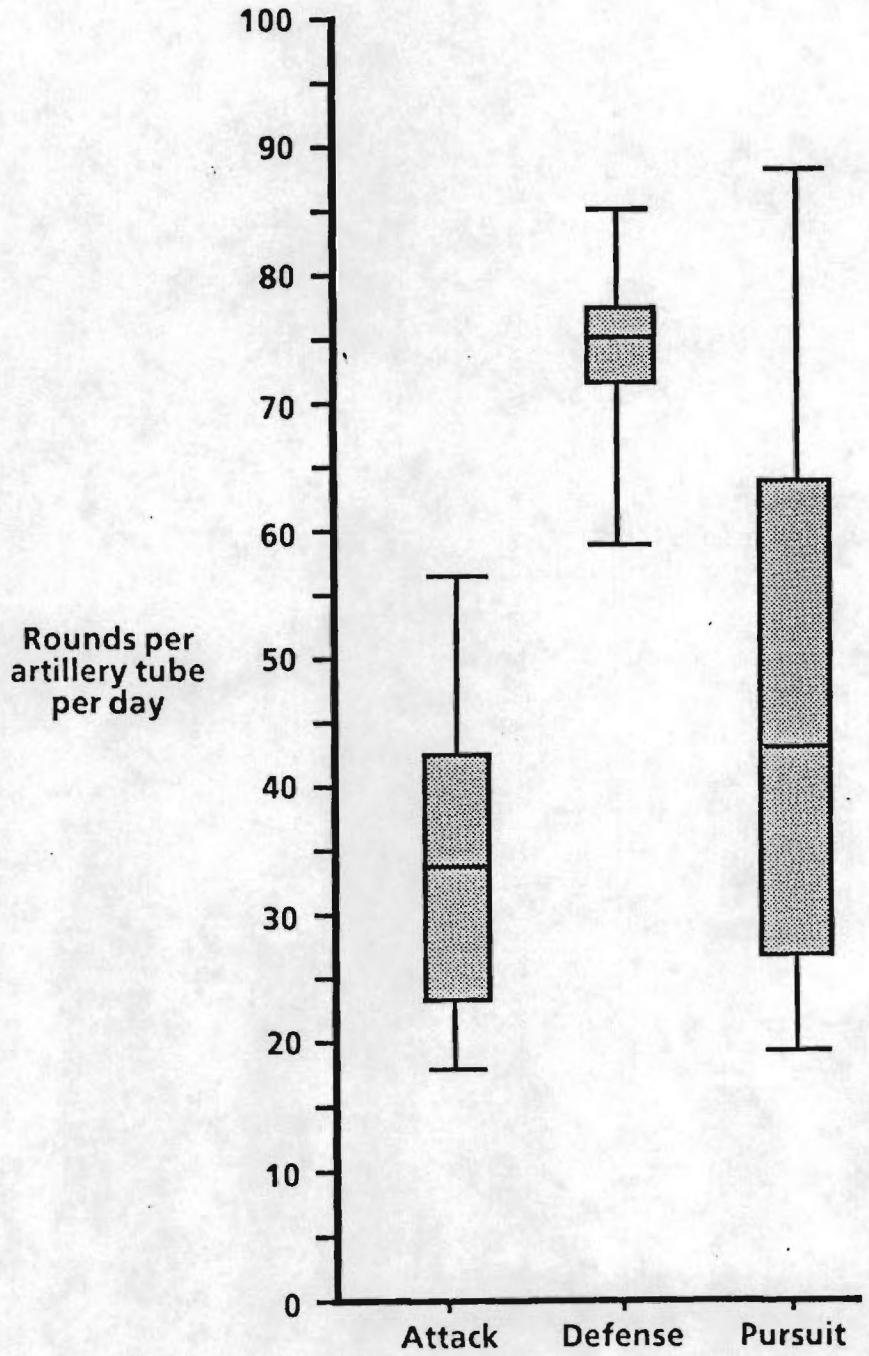


7b. TIME-DISTANCE



OUTCOMES

8. ROUNDS FIRED



Appendix D

METHODOLOGY AND VALIDATION

Proceedings of the Conference on
Methodology and Validation, 1987

6-9 April 1987
Orlando, Florida

Edited by
Osman Balci, PhD
Virginia Polytechnic Institute

Simulation Series
Volume 19
Number 1
January 1988

A Society for Computer Simulation (Simulation Councils, Inc.) publication
San Diego, California

Credibility assessment of simulation results: The state of the art

Osman Balci

Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

The purpose of this paper is to provide a state-of-the-art survey of credibility assessment of simulation results and suggest some future research directions. A hierarchy of the credibility assessment is introduced and the state-of-the-art survey is presented with respect to this hierarchy. A glossary is provided to alleviate the lack of standard terminology. The future research calls upon looking at the "global picture" when conducting a simulation study and being concerned with all of the eleven credibility assessment stages not just model validation and programmed model verification.

1. INTRODUCTION

In a report to the U.S. Congress, the U.S. General Accounting Office (U.S. GAO) [1976] reviewed 57 federally funded models in detail, each costing over $100,000 to develop, and found that many model development efforts experienced large cost overruns, prolonged delays in completion, and total user dissatisfaction with the information obtained from the model. The U.S. GAO report initiated a sequence of significant events in promoting research on model/credibility assessment.

Under the leadership of Saul I. Gass, the National Bureau of Standards organized several symposia and produced three special publications [Gass 1979, 1980, 1981]. The Society for Computer Simulation established a technical committee on model credibility which published a terminology for model credibility [Schlesinger et al. 1979]. The U.S. GAO [1979] published guidelines for model evaluation.

A uniform, standard terminology is yet nonexistent. A recent literature review [Balci and Sargent 1984a] indicated the usage of 16 terms: *acceptability, accuracy, analysis, assessment, calibration, certification, confidence, credibility, evaluation, performance, qualification, quality assurance, reliability, testing, validation,* and *verification.* Except some early papers which appeared between 1966 and 1972, model verification and model validation have been most of the time consistently defined reflecting the following differentiation:

model verification refers to building the model right; and
model validation refers to building the right model.

To alleviate the lack of standard terminology, a glossary is provided in Section 5.

The purpose of this paper is to provide a state-of-the-art survey of credibility assessment of simulation results and suggest some future research directions. A hierarchy of the credibility assessment is introduced in Section 2 and the state-of-the-art survey is presented with respect to this hierarchy in Section 3. Section 4 contains the conclusions and future research directions.

2. A HIERARCHY OF THE CREDIBILITY ASSESSMENT

To provide a proper framework for the state-of-the-art survey, it is convenient to introduce the hierarchy of the credibility assessment of simulation results as depicted in Figure 1 [Balci 1986]. Each branch of the hierarchy represents a credibility assessment stage (CAS) or an indicator. Figure 1 reveals the effect of a CAS upon the other. For example, model validity can be assessed in terms of several indicators each being a subjective or an objective test. Model validity affects the quality of experimental model which in turn affects the credibility of simulation results.

There are two more CASs not shown in Figure 1: presentation verification and acceptability of simulation results (see [Balci 1986] for details). The credibility assessment and presentation verification affect the acceptability of simulation results.

3. THE STATE OF THE ART

Recently, Banks et al. [1986a,1986b,1987] provided an excellent overview of modeling processes, validation, and verification and proposed a methodology. Gass [1983], in his feature article, presented an excellent review of the issues related to the credibility assessment. Ören [1981] proposed a frame of reference for the concepts and criteria to assess acceptability of simulation studies.

3.1 Formulated Problem Verification

Problem formulation and its verification which greatly affect the credibility and acceptability of simulation results, have not received the attention that they deserve in a simulation study. This is an educational problem. Educators usually emphasize how to solve a given problem rather than how to formulate one. As a result, people tend to jump into the solution of the communicated problem without spending sufficient time and effort in formulating the *real* problem. The consequence of this practice is frequently the type III error.

Balci and Nance [1985] introduced the formulated problem verification as an explicit requirement of model credibility. They provided a high-level procedure for problem formulation and proposed 38 indicators for evaluating a formulated problem.

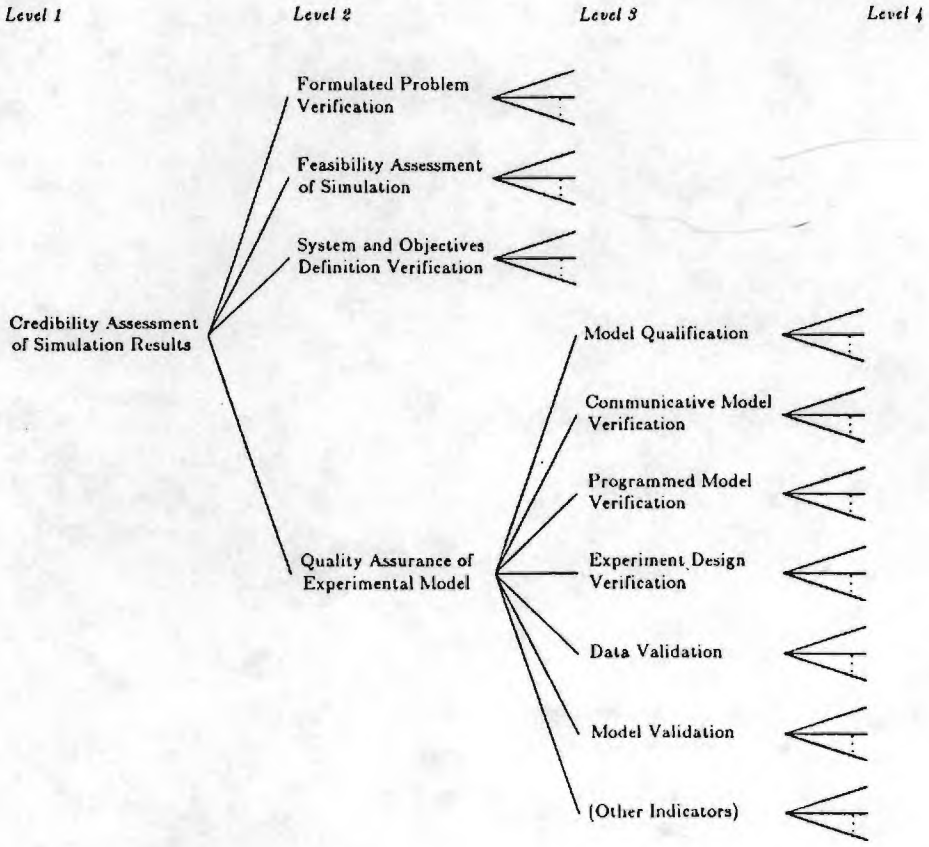


Figure 1. A Hierarchy of the Credibility Assessment.

3.2 Feasibility Assessment of Simulation

It is well to remember the dictum that if a hammer is the only tool you have, you may tend to view each problem as a nail. One should not jump into simulation without assessing its feasibility for solving the problem under study. On the other hand, the statement "when all else fails, use simulation" is misleading if not invalid. Another technique may provide a less costly solution, but it may not be as useful. See [Balci 1986] for some indicators of the feasibility of simulation.

3.3 System and Objectives Definition Verification

The system of concern here is the one which contains the whole formulated problem. Although study objectives are specified within the formulated problem, it is extremely important to explicitly define and verify them since the rest of the simulation study are based upon those objectives. System definition should be verified in terms of the system characteristics identified by Shannon [1975]: (1) change, (2) environment, (3) counterintuitive behavior [Forrester 1971], (4) drift to low performance, (5) interdependency, and (6) organization. None of the energy models could predict the oil embargo in 1973; because, at the time, it was a counterintuitive behavior. Incorrect identification of system characteristics may result in type II or type III error.

3.4 Model Qualification

A model, by definition, is an abstraction of the reality. Many assumptions are made with respect to the study objectives in abstracting the reality (system). These assumptions define the underpinnings of the model and their reasonableness must be assessed as early as possible in the model development life cycle. Using a model without knowing or understanding its underlying assumptions is absurd.

Model qualification has been studied by Gass and Thompson [1980] under the name of *theoretical validity* and by Sargent [1985] under the name of *conceptual model validity*.

3.5 Communicative Model Verification

How well the communicative model can be verified is dependent upon how much its form of representation lends itself to formal analysis and verification. Balci [1986] identified 21 forms of representation suggested in the literature.

Nance and Overstreet [1986] proposed several diagnostics which are based on analysis of graphs constructed from

a particular form of model specification called condition specification [Overstreet 1982; Overstreet and Nance 1985]. *Data-Flow Analysis* and *Control-Flow Analysis* [Adrion et al. 1982] are the other two graph-based analysis techniques applicable for communicative model verification. *Desk Checking* [Adrion et al. 1982] and *Model Review* [Balci 1986] are also useful.

3.6 Programmed Model Verification

Graph-based analysis, desk checking, and model review can also be used for the verification of a programmed model. In addition, Balci [1986] proposed the use of *Instrumentation-Based Testing* and *Functional Testing.*

3.7 Experiment Design Verification

Since all simulation models are descriptive, it is the responsibility of the simulation analyst to correctly interpret the model results. To aid the analyst in this interpretation, experiments are designed and incorporated into the programmed model producing the experimental model with which the experiments are conducted and results are obtained. Incorrect design of experiments may result in inaccurate interpretation of model results.

It is well to remember the dilemma of the scientific method as pointed out by Blyth [1973]: "The scientist needs to be objective, but the way he [or she] makes progress is through following up subjective insights." When a statistical procedure is used, we think that we are using an objective method. When it comes to satisfying the assumptions underlying the procedure, however, we sometimes use our subjective insights, intuitions, and guesses.

Balci [1986] proposed some indicators for verifying the design of simulation experiments.

3.8 Data Validation

U.S. GAO [1979] proposed a two-step approach for data validation: (1) establish the accuracy, completeness, impartiality, and appropriateness of the original data, and (2) verify the manner in which the model deals with the transformation of the original data. U.S. GAO [1979] also provided some indicators for data validity. Emphasizing the validation of input data models, Balci [1986] proposed some indicators as well.

3.9 Model Validation

The existing literature on simulation model validation [Balci and Sargent 1984a] generally falls into two broad areas: subjective validation techniques and statistical techniques proposed for validation. Tables 1 and 2 list these techniques and contain the related reference(s). The applicability of the techniques in Tables 1 and 2 depends upon the following cases where the system being modeled is: (1) completely observable—all data required for validation can be collected from system, (2) partially observable—some required data can be collected, and (3) nonexistent or completely unobservable. The statistical techniques in Table 2 are applicable only for case 1.

3.10 Quality Assurance of Experimental Model

The quality of experimental model is assured by way of integrating the six CASs and other indicators shown in Figure 1. The other indicators are given by Balci [1986] as follows: accessibility, accountability, accuracy, augmentability, communicativeness, completeness, conciseness, consistency, device-independence, efficiency, legibility, self-containedness, self-descriptiveness, structuredness, and robustness.

3.11 Credibility Assessment of Simulation Results

The credibility of simulation results is assessed by way of integrating the following four CASs: formulated problem verification, feasibility assessment of simulation, system and objectives definition verification, and quality assurance of experimental model.

4. CONCLUSIONS AND RESEARCH DIRECTIONS

As illustrated by the survey, most work has concentrated on model validation and very little has been published on the other ten CASs. However, as indicated by the hierarchy in Figure 1, model validity is a necessary but not a sufficient requirement for the credibility of simulation results. Future research should concentrate on all of the CASs.

Subjectivity is and will always be part of the credibility assessment for a reasonably complex simulation study. The reason for subjectivity is two-fold: modeling is an art and credibility assessment is situation dependent. The approach using the concept of indicators proposed by Balci [1986] is promising; however, future research is needed to determine more indicators for the CASs especially for specific areas of application (e.g., combat system simulation, manufacturing system simulation, missile system simulation, etc.).

We apparently lack good quality education on the art of modeling. It is not uncommon to find people who use the results of a simulation model without any idea about the underlying model assumptions. The dictum stated by Elmaghraby [1968] has not been fully appreciated: "Nobody solves *the* problem. Rather, everybody solves the model that he [or she] has constructed of the problem."

5. GLOSSARY

Calibration. An iterative process in which a probabilistic characterization for an input variable or a fixed value for a parameter is tried until the model is found to be sufficiently valid.

Communicative Model. A model representation which can be communicated to other humans and can be judged or compared against the system and the study objectives by more than one human [Nance 1981].

Table 1. Subjective Validation Techniques.

Event Validation ..... [Hermann 1967]

Face Validation ..... [Hermann 1967]

Field Tests ..... [Shannon 1975; Van Horn 1971]

Graphical Comparisons ..... [Cyert 1966; Forrester 1961; Miller 1975; Wright 1972]

Historical Methods ..... [Naylor and Finger 1967]

Hypothesis Validation ..... [Hermann 1967]

Internal Validation ..... [Hermann 1967]

Multistage Validation ..... [Naylor and Finger 1967; Law and Kelton 1982]

Predictive Validation ..... [Emshoff and Sisson 1970]

Schellenberger's Criteria ..... [Schellenberger 1974; U.S. General Accounting Office 1979]

Sensitivity Analysis ..... [Hermann 1967; Miller 1974a, 1974b; Van Horn 1971; Shannon 1975]

Submodel Testing ..... [Balci 1981]

Turing Test ..... [Mitroff 1969; Schruben 1980; Turing 1963; Van Horn 1971]

Table 2. Statistical Techniques Proposed for Validation.

Analysis of Variance ..... [Naylor and Finger 1967]

Confidence Intervals/Regions ..... [Balci and Sargent 1984b; Law and Kelton 1982; Shannon 1975]

Factor Analysis ..... [Cohen and Cyert 1961]

Hotelling's T^2 Tests ..... [Balci and Sargent 1981, 1982a, 1982b, 1983; Shannon 1975]

Multivariate Analysis of Variance ..... [Garratt 1974]

- — Standard MANOVA
- — Permutation Methods
- — Nonparametric Ranking Methods

Nonparametric Goodness-of-fit Tests ..... [Gafarian and Walsh 1969; Naylor and Finger 1967]

- — Kolmogorov-Smirnov Test
- — Cramer-Von Mises Test
- — Chi-square Test

Nonparametric Tests of Means ..... [Shannon 1975]

- — Mann-Whitney-Wilcoxon Test
- — Analysis of Paired Observations

Regression Analysis ..... [Aigner 1972; Cohen and Cyert 1961; Howrey and Kelejian 1969]

Theil's Inequality Coefficient ..... [Kheir and Holmes 1978; Rowland and Holmes 1978; Theil 1961]

Time Series Analysis

- — Spectral Analysis ..... [Fishman and Kiviat 1967; Gallant et al. 1974; Howrey and Kelejian 1969; Hunt 1970; Van Horn 1971; Watts 1969]
- — Correlation Analysis ..... [Watts 1969]
- — Error Analysis ..... [Damborg and Fuller 1976; Tytula 1978]

t-Test ..... [Shannon 1975; Teorey 1975]

Communicative Model Verification. Ensuring that the communicative model is correctly constructed as intended and confirming the adequacy of the communicative model to provide an acceptable level of agreement for the domain of intended application.

Conceptual Model. The model which is formulated in the mind of the modeler [Nance 1981].

Data Validation. Substantiating that each input data model used possesses satisfactory accuracy consistent with the study objectives and confirming that the simulation model parameter values are accurately identified and used.

Descriptive Model. A model which describes the behavior of a system without any value judgment on the "goodness" or "badness" of such behavior [Elmaghraby 1968].

Domain of Applicability. The set of prescribed conditions for which the experimental model has been tested, compared against the system to the extent possible, and judged suitable for use [Schlesinger et al. 1979].

Domain of Intended Application. The prescribed conditions for which the model is intended to match the system under study [Schlesinger et al. 1979].

Experiment Design. The process of formulating a plan to gather the desired information at minimal cost and to enable the analyst to draw valid inferences [Shannon 1975].

Experiment Design Verification. Substantiating that the experiments are correctly designed as intended.

Experimental Model. The programmed model incorporating an executable description of an experiment design.

Formulated Problem Verification. Substantiating that the formulated problem contains the *actual* problem in its entirety and is sufficiently well structured to permit the derivation of a sufficiently credible solution.

Indicator. An indirect measure of a concept, that can be measured directly.

Level of Agreement. The required correspondence between the model and the system, consistent with the domain of intended application and the study objectives [Schlesinger et al. 1979].

Model Builder's Risk. The probability of committing type I error.

Model Certification. Confirmation (usually by a third party) that a simulation model, within its domain of applicability, can produce results which are sufficiently credible with respect to the study objectives.

Model Qualification. Justifying that all assumptions underlying the conceptual model are appropriate and the conceptual model provides an adequate representation of the system under study with respect to the study objectives.

Model User's Risk. The probability of committing type II error.

Model Validation. Substantiating that the experimental model, within its domain of applicability, behaves with satisfactory accuracy consistent with the study objectives.

Peer Assessment. The assessment of the acceptability/credibility of simulation results by a panel of expert peers.

Prescriptive Model. A model which describes the behavior of a system with a value judgment on the "goodness" or "badness" of such behavior [Elmaghraby 1968].

Programmed Model. A model representation that admits execution by a computer to produce results [Nance 1981].

Programmed Model Verification. Substantiating that the programmed model represents the communicative model within an acceptable range of accuracy consistent with the study objectives.

System and Objectives Definition Verification. Substantiating that the system characteristics are correctly identified and the study objectives are explicitly defined with sufficient accuracy.

Type I Error. The error of rejecting the results of a simulation study when in fact they are sufficiently credible.

Type II Error. The error of accepting the results of a simulation study when in fact they are *not* sufficiently credible.

Type III Error. The error of solving the wrong problem.

ACKNOWLEDGMENTS

This research was sponsored in part by the Naval Sea Systems Command and the Office of Naval Research under Contract N60921-83-G-A165 through the Systems Research Center at VPI&SU.

REFERENCES

Adrion, W.R., M.A. Branstad, and J.C. Cherniavsky (1982), "Validation, Verification, and Testing of Computer Software," *Computing Surveys 14*, 2 (June), 159-192.

Aigner, D.J. (1972), "A Note on Verification of Computer Simulation Models," *Management Science 18*, 11 (Nov.), 615-619.

Balci, O. (1981), "Statistical Validation of Multivariate Response Simulation Models," Ph.D. Dissertation, Syracuse University, Syracuse, N.Y., Aug.

Balci, O. (1986), "Guidelines for Successful Simulation Studies: Part I and II", Technical Report TR-85-2, Department of Computer Science, Virginia Tech, Blacksburg, Va., Sept.

Balci, O. and R.E. Nance (1985), "Formulated Problem Verification as an Explicit Requirement of Model Credibility," *Simulation 45*, 2 (Aug.), 76-86.

Balci, O. and R.G. Sargent (1981), "A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models," *Communications of the ACM 24*, 4 (Apr.), 190-197.

Balci, O. and R.G. Sargent (1982a), "Some Examples of Simulation Model Validation Using Hypothesis Testing," In *Proceedings of the 1982 Winter Simulation Conference* (San Diego, Calif., Dec. 6-8). IEEE, New Jersey, pp. 620-629.

Balci, O. and R.G. Sargent (1982b), "Validation of Multivariate Response Models Using Hotelling's Two-Sample T^2 Test," *Simulation 39*, 6 (Dec.), 185-192.

Balci, O. and R.G. Sargent (1983), "Validation of Multivariate Response Trace-Driven Simulation Models," In *Performance '83*, A.K. Agrawala and S.K. Tripathi, Eds. North-Holland Publ., Amsterdam, pp. 309-323.

Balci, O. and R.G. Sargent (1984a), "A Bibliography on the Credibility Assessment and Validation of Simulation and Mathematical Models," *Simuletter 15*, 3 (July), 15-27.

Balci, O. and R.G. Sargent (1984b), "Validation of Simulation Models via Simultaneous Confidence Intervals," *American Journal of Mathematical and Management Sciences 4*, 3&4, 375-406.

Banks, J., D.M. Gerstein, and S.P. Searles (1986a), "The Verification and Validation of Simulation Models: A Methodology," Technical Report, School of Industrial and Systems Engineering, Georgia Tech, Atlanta, Ga., Sept.

Banks, J., D.M. Gerstein, and S.P. Searles (1986b), "The Verification and Validation of Simulation Models: Unresolved Issues," Technical Report, School of Industrial and Systems Engineering, Georgia Tech, Atlanta, Ga., Oct.

Banks, J., D.M. Gerstein, and S.P. Searles (1987), "Modeling Processes, Validation, and Verification of Complex Simulations: A Survey," In *Proceedings of the Conference on Simulation Methodology and Validation* (Orlando, Fla., Apr. 6-9). SCS, San Diego, Calif.

Blyth, C.R. (1973), "Subjective vs. Objective Methods in Statistics," *The American Statistician 26*, 3 (June), 20-22.

Cohen, K.J. and R.M. Cyert (1961), "Computer Models in Dynamic Economics," *Quarterly Journal of Economics 75*, 1 (Feb.), 112-127.

Cyert, R.M. (1966), "A Description and Evaluation of Some Firm Simulations," In *Proceedings of the IBM Scientific Computing Symposium on Simulation Models and Gaming* (White Plains, N.Y.), IBM, White Plains, N.Y., pp. 3-22.

Damborg, M.J. and L.F. Fuller (1976), "Model Validation Using Time and Frequency Domain Error Measures," ERDA Report 76-152, available from NTIS, Springfield, Va.

Elmaghraby, S.E. (1968), "The Role of Modeling in IE Design," *Industrial Engineering 19*, 6 (June), 292-305.

Emshoff, J.R. and R.L. Sisson (1970), *Design and Use of Computer Simulation Models*, MacMillan, New York.

Fishman, G.S. and P.J. Kiviat (1967), "The Analysis of Simulation Generated Time Series," *Management Science 13*, 7 (July), 525-557.

Forrester, J.W. (1961), *Industrial Dynamics*. MIT Press, Cambridge, Mass.

Forrester, J.W. (1971), "Counterintuitive Behavior of Social Systems," *Technology Review 73*, 3 (Jan.), 1-16.

Gafarian, A.V. and J.E. Walsh (1969), "Statistical Approach for Validating Simulation Models by Comparison with Operational Systems," In *Proceedings of the 4th International Conference on Operations Research*, John Wiley & Sons, New York, pp. 702-705.

Gallant, A.R., T.M. Gerig, and J.W. Evans (1974), "Time Series Realizations Obtained According to an Experimental Design," *J. American Statistical Association 69*, 347 (Sept.), 639-645.

Garratt, M. (1974), "Statistical Validation of Simulation Models," In *Proceedings of the 1974 Summer Computer Simulation Conference* (Houston, Tex., July 9-11). Simulation Councils, La Jolla, Calif., pp. 915-926.

Gass, S.I., Ed. (1979), *Utility and Use of Large-Scale Mathematical Models*, Special Publication 534, Nat. Bur. of Standards, Washington, D.C.

Gass, S.I., Ed. (1980), *Validation and Assessment Issues of Energy Models*, Special Publication 569, Nat. Bur. of Standards, Washington, D.C., Feb.

Gass, S.I., Ed. (1981), *Validation and Assessment of Energy Models*, Special Publication 616, Nat. Bur. of Standards, Washington, D.C., Oct.

Gass, S.I. (1983), "Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis," *Operations Research 31*, 4 (July-Aug.), 603-631.

Gass, S.I. and B.W. Thompson (1980), "Guidelines for Model Evaluation: An Abridged Version of the U.S. General Accounting Office Exposure Draft," *Operations Research 28*, 2 (Mar.-Apr.), 431-439.

Hermann, C.F. (1967), "Validation Problems in Games and Simulations with Special Reference to Models of International Politics," *Behavioral Science 12*, 3 (May), 216-231.

Howrey, P. and H.H. Kelejian (1969), "Simulation Versus Analytical Solutions," In *The Design of Computer Simulation Experiments*, T.H. Naylor, Ed. Duke University Press, Durham, N.C., pp. 207-231.

Hunt, A.W. (1970), "Statistical Evaluation and Verification of Digital Simulation Models Through Spectral Analysis," Ph.D. Dissertation, The University of Texas at Austin, Austin, Tex.

Kheir, N.A. and W.M. Holmes (1978), "On Validating Simulation Models of Missile Systems," *Simulation 30*, 4 (Apr.), 117-128.

Law, A.M. and W.D. Kelton (1982), *Simulation Modeling and Analysis*, McGraw-Hill, New York.

Miller, D.K. (1975), "Validation of Computer Simulations in the Social Sciences," In *Proceedings of the Sixth Annual Conference on Modeling and Simulation* (Pittsburg, Pa.), pp. 743-746.

Miller, D.R. (1974a), "Model Validation Through Sensitivity Analysis," In *Proceedings of the 1974 Summer Computer Simulation Conference* (Houston, Tex., July 9-11). Simulation Councils, La Jolla, Calif., pp. 911-914.

Miller, D.R. (1974b), "Sensitivity Analysis and Validation of Simulation Models," *J. Theoretical Biology 48*, 2 (Dec.), 345-360.

Mitroff, I.I. (1969), "Fundamental Issues in the Simulation of Human Behavior: A Case in the Strategy of Behavioral Science," *Management Science 15*, 12 (Dec.), 635-649.

Nance, R.E. (1981), "Model Representation in Discrete Event Simulation: The Conical Methodology," Technical Report CS81003-R, Department of Computer Science, Virginia Tech, Blacksburg, Va., Mar.

Nance, R.E. and C.M. Overstreet (1986), "Diagnostic Assistance Using Digraph Representations of Discrete Event Simulation Model Specifications," Technical Report TR-86-8, Department of Computer Science, Virginia Tech, Blacksburg, Va., Mar.

Naylor, T.H. and J.M. Finger (1967), "Verification of Computer Simulation Models," *Management Science 14*, 2 (Feb.), B92-B101.

Ören, T.I. (1981), "Concepts and Criteria to Assess Acceptability of Simulation Studies: A Frame of Reference," *Communications of the ACM 24*, 4 (Apr.), 180-189.

Overstreet, C.M. (1982), "Model Specification and Analysis for Discrete Event Simulation," Ph.D. Dissertation, Virginia Tech, Blacksburg, Va., Dec.

Overstreet, C.M. and R.E. Nance (1985), "A Specification Language to Assist in Analysis of Discrete Event Simulation Models," *Communications of the ACM 28*, 2 (Feb.), 190-201.

Rowland, J.R. and W.M. Holmes (1978), "Simulation Validation with Sparse Random Data," *Computers and Electrical Engineering 5*, 3 (Mar.), 37-40.

Sargent, R.G. (1985), "An Expository on Verification and Validation of Simulation Models," In *Proceedings of the 1985 Winter Simulation Conference* (San Francisco, Calif., Dec. 11-13). IEEE, Piscataway, N.J., pp. 15-22.

Schellenberger, R.E. (1974), "Criteria for Assessing Model Validity for Managerial Purposes," *Decision Sciences 5*, 4 (Apr.), 644-653.

Schlesinger, S., et al. (1979), "Terminology for Model Credibility," *Simulation 32*, 3 (Mar.), 103-104.

Schruben, L.W. (1980), "Establishing the Credibility of Simulations," *Simulation 34*, 3 (Mar.), 101-105.

Shannon, R.E. (1975), *Systems Simulation: The Art and Science*, Prentice-Hall, Englewood Cliffs, N.J.

Teorey, T.J. (1975), "Validation Criteria for Computer System Simulations," *Simuletter 6*, 4 (July), 9-20.

Theil, H. (1961), *Economic Forecasts and Policy*, North-Holland Publ., Amsterdam.

Turing, A.M. (1963), "Computing Machinery and Intelligence," In *Computers and Thought*, E.A. Feigenbaum and J. Feldman, Eds. McGraw-Hill, New York, pp. 11-15.

Tytula, T.P. (1978), "A Method for Validating Missile System Simulation Models," Technical Report E-78-11, U.S. Army Missile R&D Command, Redstone Arsenal, Ala., June.

U.S. General Accounting Office (1976), "Report to the Congress: Ways to Improve Management of Federally Funded Computerized Models," LCD-75-111, U.S. G.A.O., Washington, D.C., Aug.

U.S. General Accounting Office (1979), "Guidelines for Model Evaluation," PAD-79-17, U.S. G.A.O., Washington, D.C., Jan.

Van Horn, R.L. (1971), "Validation of Simulation Results," *Management Science 17*, 5 (May), 247-258.

Watts, D. (1969), "Time Series Analysis," In *The Design of Computer Simulation Experiments*, T.H. Naylor, Ed. Duke University Press, Durham, N.C., pp. 165-179.

Wright, R.D. (1972), "Validating Dynamic Models: An Evaluation of Tests of Predictive Power," In *Proceedings of the 1972 Summer Computer Simulation Conference* (San Diego, Calif., July 14-16). Simulation Councils, La Jolla,